A Systematic Literature Review of Data Science, Data Analytics and Machine Learning Applied to Healthcare Engineering Systems

Abstract

Purpose – The objective of this paper is to assess and synthesize the published literature related to the application of data analytics, big data, data mining, and machine learning to healthcare engineering systems.

Design/methodology/approach – A systematic literature review (SLR) was conducted to obtain the most relevant papers related to the research study from three different platforms: EBSCOhost, ProQuest, and Scopus. The literature was assessed and synthesized, conducting analysis associated with the publications, authors, and content.

Findings – From the SLR, 576 publications were identified and analyzed. The research area seems to show the characteristics of a growing field with new research areas evolving and applications being explored. In addition, the main authors and collaboration groups publishing in this research area were identified throughout a social network analysis. This could lead new and current authors to identify researchers with common interests on the field.

Research limitations/implications – The use of the SLR methodology does not guarantee that all relevant publications related to the research are covered and analyzed. However, the authors’ previous knowledge and the nature of the publications were used to select different platforms.

Originality/value – To the best of the authors’ knowledge, this paper represents the most comprehensive literature-based study on the fields of data analytics, big data, data mining, and machine learning applied to healthcare engineering systems.
Keywords: Data analytics, big data, machine learning, healthcare systems, systematic literature review

Article classification: Literature review

1. Introduction

Data science is a “set of fundamental principles that support and guide the principled extraction of information and knowledge from data” (Provost and Fawcett, 2013). It involves the use and development of algorithms, processes, methodologies, and techniques for understanding past, present, and future phenomena through the analysis of data to improve decision-making. Data scientists and data analytics must be able to view business problems from a data perspective to be able to leverage the benefits of its application on the organization. The healthcare industry is one of the world’s largest, most critical, and fastest-growing industries that is evolving through significant challenges in recent times (Nambiar et al., 2013). It is considered as a data-driven industry and has historically generated a large amount of data, driven by record keeping, compliance and regulatory requirements, and patient care (Raghupathi and Raghupathi, 2014). However, according to a report from the Institute of Medicine, the healthcare industry is considered a highly inefficient industry, where one-third of its expenditures are wasted and do not contribute to better quality outcomes. While the healthcare system continues to apply industrial and systems engineering tools to achieve an effective coordinated system, data analytics have the potential to improve care, save lives and lower costs by identifying associations and understanding trends and patterns within the data.

Data science has several areas and disciplines within itself; thus, there is no universal agreement in the literature regarding its components and interactions. Winters (2015) developed a Venn diagram to visualize the three main fields of data science (i.e., data analytics, big data, and
algorithms) and their intersections (i.e., data mining, machine learning, and software tools) based on a two-axis diagram (i.e., on the x-axis: experimental versus theoretical; on the y-axis: descriptive versus prescriptive). On the other hand, Emmert-Streib et al., (2016) developed a schematic visualization (i.e., Efron-triangle) of the main fields constituting data science (i.e., domain knowledge, statistics/mathematics, computer science) and their intersections (i.e., machine learning, biostatistics, data engineering), based on the original data science Veen diagram created by Conway (2013). Taking into consideration the significant role data science can take to achieve better outcomes in healthcare systems, it would be relevant to understand to what extent each field/area has been applied, and its maturity state, in healthcare systems, along with the authors researching that field/area. Therefore, the purpose of this study is to assess and synthesize the published literature related to the impact, benefits, implications, challenges, opportunities, and trends of data science exclusively in healthcare systems. To achieve this aim, the authors used a SLR as the research methodology. SLRs focus on the published literature of a specific research field by identifying, evaluating, and integrating the findings of all relevant studies that address a set of research questions while being objective, systematic, transparent, and replicable. However, for highly relevant publications to be identifiable, they must be indexed in targeted platforms/databases (Lefebvre et al., 2011). To ensure this, the authors have strategically selected platforms that contained medical databases to provide adequate coverage of the research area and designed a search strategy that allowed the capture of as many significant publications as possible.

After the final set of publications was obtained for this study, three different dimensions were assessed and evaluated to synthesize information, i.e., publication characteristics, authors’ characteristics, and content characteristics. These were identified based on preliminary work that
defined relevant criteria to assess the maturity of a research area (Keathley et al., 2013). The publication characteristics analyses included an examination of the publication trends over time as well as the characteristics of the publications’ sources associated with the final paper set, which in this case were primarily academic journals, given the nature of the publication set. The authors’ characteristics examination included an investigation of author quantities and author collaborations among them through social network analyses to identify predominant authors and research groups. Investigation of content characteristics, for this work’s purpose, refers to analyze the scope in which the areas/fields within data science (e.g., data analytics, machine learning, data mining) have been addressed in healthcare systems, in which medical areas/departments and to treat which diseases/disorders. To address this, a social network analysis was conducted. Thus, the research questions addressed in this study are:

a. Publications characteristics: Which trend exists in publication pattern overtime for this research area (RQ1)?, what type of sources are publishing the works (RQ2)?, which are the sources with the highest frequency of published works on the field (RQ3)?, and which are the main study fields from the sources publishing the works (RQ4)?

b. Authors’ characteristics: How many authors are contributing to this area (RQ5)?, to what extent are new authors contributing (RQ6)?, to what extent are authors collaborating among them in this research area (RQ7)?, and what is the distribution number of authors per publication (RQ8)?

c. Content characteristics: Which are the most frequently mentioned data science fields applied to healthcare systems (RQ9)?, which are the top medical areas/departments where data science has been studied and applied (RQ10)?, which are the top diseases/disorders being addressed through data science approaches (RQ11)?, which are the main study
approaches on the theoretical publications set (RQ12)?, which are the main application objectives on the case study publications set (RQ13)?, and which are the newly emerging research lines related to this research area (RQ14)?

The rest of the paper is divided into three main sections: the research methodology (i.e., SLR conduction) is presented in Section 2; the results of the study (publication characteristics, authors’ characteristics, and content characteristics) are included in Section 3 and; Section 4 presents the conclusions and future research directions.

2. Research methodology

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) is a well-recognized research methodology in the medical field; it uses four steps (Moher et al., 2010), namely: identification, screening, eligibility, and included. This research method is often used in meta-analyses. On the other hand, the systematic literature review approach proposed by Keathley et al., (2016), based on Tranfield et al., (2003) and the Cochrane Handbook (Higgins and Green, 2011; Lefebvre et al., 2011), has been used in bibliometric and/or scientometric analyses. Keathley et al., (2016) follows seven steps:

a) Problem definition: the research area is identified and the research objectives defined.

b) Scoping study: the desired scope of the study is established and the research team conducts a ‘traditional’ literature review to identify relevant publications related to the research area.

c) Search strategy: the scoping set of papers is evaluated by identifying potential search terms. Then, the strategy is formulated by defining the databases/platforms to be searched, Boolean phrases, search tools, limiters, filters, and exclusion criteria.
d) **Exclusion criteria:** Publications not directly related to analytics, data mining, big data, and machine learning applied in healthcare engineering systems are excluded.

e) **Data collection:** bibliometric data is collected and the criteria identified based on the aim of the research study.

f) **Data analysis:** the bibliometric analysis is conducted based on the aim of the research study.

g) **Reporting:** findings and results are presented.

In this study, the research team decided to use Keathley et al., (2016) research methodology based on two considerations. First, the purpose of this study was focused on conducting quantitative analyses of published documents, also known as bibliometric analyses (Broadus, 1987). Second, Keathley et al., (2016) included in their research methodology three critical steps (problem definition, scoping study, and search strategy) that are not included in PRISMA. These three steps offer the possibility of easily updating a systematic literature review.

**2.1 Problem definition**

Throughout the literature, there are multiple publications regarding the use of data science, data analytics, and machine learning algorithms applied to healthcare systems. However, it is not clear to what extent authors contributing to this research area are collaborating to diffuse new knowledge and significant findings. For this reason, a SLR aiming to synthesize the current published literature would provide a guide for the future development and evolution of this research area.

**2.2 Scoping study**

The scoping study was conducted through the identification of six main publications related to the research area using three platforms (EBSCOhost, ProQuest, and Scopus): Malik *et al.* (2018),...
Islam et al. (2018), Hansen et al. (2014), Luo et al. (2016), Alonso et al. (2017), and Mehta and Pandit (2018). To determine to what extent the literature related to data science applied to healthcare systems had been analyzed, a comparison study of previous literature reviews was conducted (see Table 1). The literature review conducted in 2014 aimed to discuss the perspectives of the evolving use of big data in science and healthcare and to examine some of the opportunities and challenges. The literature review conducted in 2015 discussed big data applications in four majors biomedical subdisciplines: bioinformatics, clinical informatics, imaging informatics, and public health informatics. The literature review carried out in 2017 reviewed big data sources and techniques in the health sector and identified which of these techniques were the most used in the prediction of chronic diseases. Once again, the first literature review conducted in 2018 reviewed big data analytics applications and challenges in its adoptions in healthcare, and identified strategies to overcome them. The second literature review conducted in 2018, the most extensive one, provided a systematic review of the development of the fields of multiple healthcare sub-areas, data mining techniques, types of analytics, data and data sources, as well as possible directions. Finally, the last literature review conducted in 2018 assessed and synthesized how the big data phenomenon has contributed to better outcomes for the delivery of healthcare services.

One interesting finding from these systematic literature reviews is the fact that none of them conducted social network analyses related to authors publishing in this research field, which represented a gap within this field to be covered. The present study, in addition to being the most updated one, analyzed a significantly higher number of publications in comparison with these other studies. Including a theoretical approach study as well as a social network of the authors publishing in the research field aiming to help new and current researches identify researchers
who have similar interests and research lines within this field and that are collaborating in study
groups for the diffusion of knowledge.

2.3 Search strategy

The initial search strategy protocol consisted of 5 single search terms (data analytics, big data,
data mining, machine learning, and healthcare), three platforms (EBSCOhost, ProQuest, and Scopus),
the utilization of Boolean operators (AND/OR), all fields search and two main exclusion
criteria—published in academic journals and written in the English language. This
search strategy was tested and modified multiple times to identify a final set of relevant
publications for this research area. First, to increase the sensitivity of the search, synonyms
(e.g., data analysis, analysis of data, mass data, massive data), techniques (e.g., data
processing, text mining, deep learning), more specific concepts (e.g., artificial intelligence,
business intelligence, internet of things), and the term “health care” (due to the lack of
standardization between healthcare and health care in publications and academic texts) were
added into the original search terms using the OR Boolean operator. Second, also to increase
sensitivity, the Boolean phrase was applied to abstracts instead of all fields or all text,
which helped control the scope. Lastly, conference materials were considered in the
publications’ search. Table 2 shows the final search strategy protocol used in this work.
The search strategy was executed to identify all relevant papers up through July 2019.

2.4 Exclusion criteria

A total of 8,529 publications were identified and screened based on the exclusion criteria listed
in Table 2, removing the following publications from this study: duplicated (16.4% of the raw
results), not related to data science fields (29.4%), not exclusively focused to healthcare systems (47%), and without an electronic file (0.4%). From the initial set, a total of 576 publications (6.8%) were accepted as the final publication set for this research. For purposes of this research, these 576 publications were classified into two separate sets based on their research approach: theoretical and application publications. The theoretical publication set included 105 publications that mainly focused on studying and analyzing the strengths, weaknesses, opportunities, threats, challenges, capabilities, trends, benefits, and promises of data science, data analytics, and machine learning algorithms applied to healthcare systems as a whole. On the other hand, the application publication set included 471 publications related to case studies of data science, data analytics and machine learning algorithms applied to healthcare systems that addressed a specific problem, disease, medical condition, or medical disorder.

To investigate the extent to which this research area was expanding, synthesizing, and assessing the literature in the three dimensions outlined earlier (publications characteristics, authors characteristics, and content characteristics) became a significant task. Each of these included the analysis of one or more criteria, as reported in the following section.

3. Results

To obtain a comprehensive perspective of the published literature of data science, data analytics, and machine learning applied to healthcare engineering systems, this section presents the results of the analyses conducted to address the research questions posed earlier in the Introduction section.
3.1 Publications characteristics

To answer the research questions from publication characteristics, the following data were collected and synthesized from the 576 publications: publication year, publication name, publication field, and publication impact (quartile).

a. RQ1: Which trend exists in publication pattern overtime for this research area? Trends analyses are useful for visualizing trends in the frequency of publications over time to determine the extent to which the frequency is changing. When conducting a SLR, the publication rate is one of the multiple analyses often used to evaluate publication trends. Figure 1 shows the frequency of publications per year; the following findings can be observed from it. The first paper focusing on data science, data analytics, and machine learning applied to healthcare engineering systems was published in 2004 – thus, this particular research area spans only 15 years and appears to be relatively young. Second, from 2004 to 2010, the number of publications fluctuated between zero and three; and does not seem to demonstrate an increasing trend. Third, as suggested by the cumulative frequency line, the publication trend started to increase after the year 2011, being 2016 the year with the highest number of publications (195 papers), up to date. For purposes of this analysis and considering that the publication set included papers published until the end of June 2019, the last column corresponding to the frequency of published papers in 2019 was doubled to keep consistency within the data.

b. RQ2: What type of sources is publishing the works? These publications have been published mainly in academic journals (410 publications; 71.2%) and conference
proceedings (95 publications; 16.5%). This fact suggests that practitioners and academics are conducting theoretical and applied research on this topic.

c. **RQ3: Which are the sources with the highest frequency of published works in the field?** A total of 346 publication outlets were identified from the set of 576 publications. The most frequently used were academic journals such as the *Journal of Medical Systems* (33), *PLoS One* (32), *BMC Medical Informatics and Decision Making* (13), *International Journal of Advanced Research in Computer Science* (12), *BMC Bioinformatics* (11), *Journal of Big Data* (11), *Computers in Biology and Medicine* (10), and *Journal of Medical Internet Research* (10). On the other hand, the conference proceedings authors most frequently published in were the *18th IEEE International Conference on e-Health Networking, Application and Services*, *IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies*, *2016 IEEE International Conference on Healthcare Informatics*, *2016 IEEE International Conference on Mobile Services* and *2016 6th International Conference - Cloud System and Big Data Engineering*, all with two publications each, respectively. Although this research topic is limited only to healthcare engineering systems, the descriptive analysis in RQ2 shows evidence that this research topic has been addressed from different fields.

d. **RQ4: Which are the main study fields from the sources publishing the works?** An analysis was conducted to identify the publications outlets’ main study fields, according to SJR – Scimago Journal and Country Rank, to determine which research field this topic would fit better. According to the results of the analysis, the publication outlets’ main study fields were medicine (138), health informatics (101), information systems (82), computer science applications (67), computer networks and communications (61), biochemistry, genetics and
molecular biology (55), health information management (48), electrical and electronic engineering (43), agricultural, and biological sciences (37) and hardware and architecture (34). One interesting finding is the fact that most of the publication outlets’ study fields could be associated in three main fields: health, computer science, and information systems. Finally, an analysis of the journals’ impact factor quartiles (Q1 – Higher impact to Q4 – Lower impact) was conducted to identify their ranks in their respective categories: Q1 (42%), Q2 (39%), Q3 (15%), and Q4 (4%). This result suggests that most of the journals where the authors are publishing their works are highly ranked in their respective fields of study.

Overall, from publications characteristics, it is observed that this research topic (application of data analytics, big data, data mining, and machine learning to healthcare engineering systems) is in a growing stage based on the information synthesis from the analyses conducted. In essence, the frequency of publications per year shows an increasing trend, most of the publications came from journals with high impact (Q1 and Q2), and the publications are highly centered in the medical and computer sciences fields.

### 3.2 Authors’ characteristics

To answer the research questions from authors’ characteristics, the following data were collected and synthesized from the 576 publications: authors’ names, authors’ first publication year, authors’ country of affiliation, authors’ publication network (authors publishing together), and the number of authors per publication.

a. **RQ5: How many authors are contributing to this area?** A total of 2,402 unique authors were identified from the 576 publications, for an average of 4.2 authors per publication.
b. RQ6: To what extent are new authors contributing? An analysis of the frequency of new authors publishing in this research area was conducted, as shown in Figure 2. The graph suggests an increasing trend on the number of new authors publishing in this research area, being the year 2016 the one with the highest introduction of new authors; further, the cumulative frequency seems to support the ability of this research area to attract new authors. For purposes of this analysis and considering that the publication set included papers published until the end of June 2019, the last column corresponding to the frequency of new authors in 2019 was doubled to keep consistency within the data. A criterion commonly used to analyze authors’ characteristics is of author diversity, which investigated the authors’ affiliation country. This analysis allows determining to what extent authors’ interest is concentrated primarily in a geographical region or dispersed around the world. The 2,402 unique authors on both publication sets represented a total of 51 different countries. The countries with the highest number of authors were the USA (34.6%), China (15.2%), India (7.5%), United Kingdom (6.3%), and Australia (5.8%). Other countries represented South Korea, Canada, Germany, Italy, and Spain with less than 4% each. Therefore, this research area, while attracting interest from authors around the world, representing all continents, is concentrated primarily in five countries accounting for most of the authors (69.4%).

<insert Figure 2 about here>

c. RQ7: To what extent are authors collaborating between them in this research area? Collaboration among authors was analyzed using a social network created in Gephi to visualize direct and indirect interactions among authors and study groups. There are several algorithms used to draw social networks, such as Fruchterman Reingold and
Wakita. Tsurumi. The decision about which social network algorithm to use is usually based on authors’ needs (Pajntar, 2006), e.g. time consumed to process a large amount of data and drawing characteristics. For this study, the research team decided to use the Fruchterman Reingold algorithm as it is a force-directed layout algorithm that considers the force between two nodes (Udanor et al., 2016), as we were interested in analyzing the relationship among authors. Figure 3 shows the authors’ names with color-coded. In essence, authors with a blue font appeared exclusively on the theoretical publications set, authors with a black font appeared exclusively on the case study publications set, and authors with a red font appeared on both sets. For this figure, the size of the nodes represented the number of publications per author and the width of the connecting line between nodes represented the total number of publications between two given authors. The authors with the highest number of publications were I. Dinov, Francisco Florez-Revuealta, Nuno Garcia, Ivan Pires, Nuno Pombo, and S. Spisante, with four publications each, respectively. A large number of authors that have published more than a single paper suggests that this research area represents the main research focus for multiple authors. In the same way, Figure 3 illustrates the formation of multiple study groups, which confirms that diffusion of knowledge is occurring through collaboration.

<insert Figure 3 about here>

d. **RQ8: What is the distribution of the number of authors per publication?** The analysis of the number of authors per publication was performed to get an insight into how this research field is being studied (i.e., individually or in groups). Out of the 576 results, only 53 of them (or 9.20% of the analyzed publications) were written by a single author. In contrast, the other 523 publications were written in groups between 2 and 22 authors.
The group of 3 authors has the highest frequency with 117 publications (or 20.31% of the analyzed publications). With this analysis, it can be inferred that it is most likely for authors to study this research field in groups rather than individually, which strengthens the fact that the diffusion of knowledge is occurring through collaboration.

Overall, from the authors’ characteristics, it was observed that this research topic (application of data analytics, big data, data mining, and machine learning to healthcare engineering systems) is in a growing stage based on the information synthesis from the analyses conducted. In essence, the results indicate that there is a large number of authors publishing mainly in groups, the number of new authors (see Figure 2) has an increasing trend, authors’ country of affiliation are mainly focused in the U.S and China with a widespread around 51 countries, and there are groups of authors working.

3.3 Content characteristics

To answer the research questions from content characteristics, the following data were collected and synthesized from the 576 publications: publication keywords, publication approach (theoretical or case study), publication objectives, and analyses included in the publications.

a. **RQ9: Which are the most frequently mentioned data science fields applied to healthcare systems?** A total of 1,875 keywords were collected from the 576 publications and classified in 982 unique keywords. The first 28 unique keywords (2.54%) were related to data science, such as big data (81 publications), machine learning (68 publications), and data mining (65 publications). These first 28 unique keywords represented 580 out of the 1,875 keywords (31%). On the other hand, 780 unique keywords were mentioned only one time, indicating that a wide variety of topics were addressed in the 576 publications. To identify
and analyze the top data science fields and machine learning algorithms applied to healthcare systems, as well as their concurrence relationship, a social network with the keywords from both publication sets was created using Gephi (see Figure 4). Considering that the research team was interested in understanding the relationship between two keywords (nodes), then the Fruchterman Reingold clustering algorithm was applied again. Similarly, the size of the nodes represented the keyword’s count frequency, while the width of the connecting lines between nodes represented the total number of times they appeared together in a publication. The top five data science fields applied to healthcare systems where big data, machine learning, data mining, decision support systems, and the Internet of Things. On the other hand, the top machine learning, and learning algorithms applied where cloud computing, decision tree, neural networks, Naïve Bayes classifier, support vector machines, and association rule. An interesting finding is the fact that the top machine learning algorithms applied to healthcare systems were classification and clustering algorithms, which suggests an idea of the purposes behind their applications.

b. **RQ10: Which are the top medical areas/departments where data science has been studied and applied?** Identifying the top medical areas/departments where data science and machine learning algorithms have been applied allows making inferences about the application fields’ sizes, and thus, the degree to which they have been explored. According to the frequency of keywords, 24 out of the 982 unique keywords were related to different medical areas/departments. The first keyword mentioning a medical area/department was observed in the 29th place (ontology). The most frequent medical areas/departments were ontology (7 publications), mental health (5 publications), health services (4 publications),
elderly healthcare (3 publications), epidemiology (3 publications), genomics (3 publications), behavioral health (2 publications), drug development (2 publications), genomics (2 publications), and intensive care units (2 publications).

c. **RQ11: Which are the top diseases/disorders being addressed through data science approaches?** Similarly, an analysis of the top diseases being addressed through data science and machine learning algorithms was conducted. Fifty-four unique keywords related to medical disease were collected. One interesting finding is the fact that most of the disease approached can be classified into three main groups: heart diseases (e.g., cardiovascular disease and strokes) with 12 publications, cancer (e.g., breast cancer) with 9 publications, and diabetes (e.g., diabetes type 2) with 9 publications. These diseases are all top leading causes of Americans’ deaths and disabilities and leading drivers of the United States’ $3.5 trillion in annual healthcare costs, according to the National Center for Chronic Disease Prevention and Health Promotion (2019). Other diseases and medical disorders frequently studied and addressed through data science and machine learning algorithms were HIV (4 publications), asthma (3 publications), and depression (3 publications), respectively.

d. **RQ12: Which are the main study approaches on the theoretical publications set?** Table 3 classified the publications on the theoretical set based on their research area and study/analysis performed (see Appendix A to identify the reference). As suggested previously in Figure 4 and displayed in Table 3, most of the research of the publications on the theoretical set focused on big data, which is highly correlated to the amount of data generated daily by the healthcare industry.

e. **RQ13: Which are the main application objectives on the case study publications set?** Table 4 classifies the publications on the case study application set based on their application
objective (see Appendix A to identify the reference). As suggested in Figure 4 and displayed in Table 4, the application purposes of machine learning algorithms were mainly for prediction (e.g., readmissions prediction, disease prediction, fraud prediction, adverse event prediction, medical outcomes predictions), classification (i.e., based on the patients’ treats and characteristics), and decision-making (e.g., type of surgery, drugs, and recovery process). They outlined the significant role of predictive analytics in healthcare systems.

f. RQ14: Which are the newly emerging research lines related to this research area? A qualitative study was performed on the theoretical publications set to identify the newly emerging research lines. These included (1) the creation of algorithms and big data analytics technologies to address data privacy, data security and data traceability concerns, (2) improved understanding of the ethical, societal and economic implications of applying data analytics and machine learning algorithms in healthcare organizational decision-making, (3) big data and machine learning algorithms in conjunction with evidence-based medicine practices, (4) integration of multiple databases with different data structures, (5) big data applied into molecular-level data (i.e., the atomic scale), (6) applications related to social media investigation, (7) addressing information loss in data preprocessing and cleaning steps, and (8) data analysis and automation for non-experts.

<insert Table 3 about here>

<insert Table 4 about here>

Overall, from content characteristics, it was observed that this research topic (application of data analytics, big data, data mining, and machine learning to healthcare engineering systems) had been addressed from theoretical and case study approaches with a widespread of purposes.
However, from the authors’ perspective, this research topic is still in a growing stage with several medical areas/departments to study, as well as different diseases.

4. Conclusions, limitations and future research

The objective of this study was to assess and synthesized the published literature related to the application of data analytics, big data, data mining, and machine learning to healthcare engineering systems. To achieve this aim, an SLR was conducted to collect relevant publications to assess the maturity of this research field in three dimensions (Keathley et al., 2016): publication characteristics, authors’ characteristics, and content characteristics. First, the frequency of publications indicates an increasing trend, suggesting that every year more authors are publishing theoretical or application papers. Comparing Figure 1 with the life cycle of a product (introduction, growth, mature, and decline), it could be assumed that this research field is in its growth stage. These publications came from journals with a high impact factor in different fields, such as medicine, informatics, and computer science, indicating that data analytics, big data, data mining, and machine learning in healthcare engineering systems have been addressed from a multidiscipline perspective. Although these analyses are usually applied in literature reviews, the authors identified that this research topic is a skill not addressed from the industrial engineering and management decision perspective.

Second, the frequency of new authors per year supports the assumption made in the analysis of publication characteristics dimension, where it is evident that new authors are interested in this topic, contributing to the body of knowledge with their publications. On the other hand, the utilization of social network analysis was used to identify groups of authors working together to conduct theoretical and applied research in this field. Now, practitioners and academics interested in this field are able to identify them and request to participate in future investigations.
or applications of data analytics, big data, data mining, and machine learning in healthcare engineering systems. With this analysis, this paper contributes to the body of knowledge, closing a gap identified during the literature review section in this paper (see Table 1).

Lastly, the content characteristic dimension was also addressed using social network analysis to show the relationship between the keyword used in the set of publications and the classification of papers based on their study approach (theoretical research) and application objectives (applied research). The keyword social network analysis showed that this research field had been analyzed in a variety of hospital departments and illnesses. However, the authors did not find evidence of publications evaluating the impact of data analytics on patient safety and or cost versus benefits in healthcare institutions. The other two analyses included in Tables 3 and 4 showed a lack of theoretical publications focused on analyzing the decision-making process. However, from an applied research perspective, decision-making was the third most important application objective (see Table 4). On the other hand, considering the four stages of data analytics maturity (descriptive, analytic, predictive, and prescriptive) and analyzing the publications collected in the application research perspective (see Table 4), it was also observed that most of the application objectives were focused on predictive analyses. This evidence suggests that data analytics, big data, data mining, and machine learning in healthcare engineering systems had a high level of maturing. With these analyses, this paper contributes to the body of knowledge, closing a gap identified during the literature review section in this paper (see Table 1).

Using together the information shown in Figure 3, Table 3, Table 4, Appendix A, and Appendix B practitioners and academics interested in this topic should be able to easily identify new colleagues, opportunities for new research, and evidence to support the needs for specific
research. For example, application of data science and industrial engineering tools/methods to improve healthcare process efficiency, the role of data science in healthcare performance excellence models, or the creation of management decision models using data science in cases of global natural disasters or pandemics. Therefore, besides aiming at stimulating scientific research, this paper also intends to provide industrialists with a general overview of data analytics, big data, data mining, and machine learning in healthcare engineering systems so they can develop a deeper and richer knowledge on these subjects, and their practices. This will help healthcare industrialists to formulate more effective strategies for the implementation of the technologies. This research will also encourage them, and hence their organisations, to implement digital technologies to support the operations of their organisations.

These findings should not be generalized, taking into consideration that every literature review has different biases, such as database bias (produced by the utilization of a limited number of databases) and interpretation bias (produced by the interpretation of the publication content using several researchers). To reduce the impact of these biases in this research, the authors used several platforms to collect the relevant publications (ProQuest, EBSCOhost, and Scopus). Each of these platforms has access to different databases. On the other hand, under the supervision of a leading author, a single author was used to collect and interpret the information from our publication final set. Based on the current findings and the limitations of this paper, the authors consider that future research should be focused on increasing the theoretical and applied research on four lines in this field: assess cost versus benefits of the application of data analytics, conduct prescriptive analytics research, analyze decision-making process with data analytics, and update this SLR in a short time including new platforms.

References


Appendix A


Appendix B


[234] Kuo, M.H., Chrimes, D., Moa, B. and Hu, W., 2015, December. Design and construction of a big data analytics framework for health applications. In *2015 IEEE International Conference on SmartCity/SocialCom/SustainCom (SmartCity)* (pp. 631-636). IEEE.


[390] Yang, C., Delcher, C., Shenkman, E. and Ranka, S., 2016, September. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In 2016 IEEE 18th International conference on e-Health networking, applications and services (Healthcom) (pp. 1-6). IEEE.


