Data Science in Healthcare: a Systematic Review

Preprint · March 2023

DOI: 10.13140/RG.2.2.32836.19842

2 authors:

Sihem Ben Sassi
Université de la Manouba
33 PUBLICATIONS 130 CITATIONS
SEE PROFILE

Nacim Yanes
Jouf University Saudi Arabia
22 PUBLICATIONS 126 CITATIONS
SEE PROFILE

Data Science in Healthcare: a Systematic Review

Sihem Ben Sassi[a,*], Nacim Yanes[a,]

[a] *RIADI Laboratory, National School of Computer Sciences, La Manouba University, La Manouba, 2010, Tunisia*

Abstract

The digitization of data in healthcare industry has continued to increase, appealing researchers to think about data science in healthcare. There have been plenty of studies about the application of data science in healthcare fields. The objective of this work is to determine to which extent data science has been applied in healthcare and its maturity state. To reach this goal, a systematic review, guided by the three W-questions (what, why and when) was conducted. A total of 211 papers published between 2015 and 2021 were selected and studied. Results of this review outline several key directions for future research. Despite the fact that data science is well applied in various healthcare fields, particularly through machine learning techniques, our findings highlight the lack of reliable artificial wisdom techniques that allow systems to understand the overall environment and become adoptable for a wise healthcare-oriented decision.

Keywords: Data science, Data analytics, Big data, Machine learning, Healthcare, Systematic review

1. Introduction

The healthcare industry has been one of the world's largest and most fastest-growing industries that is evolving through significant challenges in recent times (El Khatib et al., 2022; OECD, 2021; Nambiar et al., 2013). It is considered as

[*] Corresponding author

Email addresses: sihem.bensassi@gmail.com (Sihem Ben Sassi), nacim.yanes@gmail.com (Nacim Yanes)

a data-driven industry and has historically generated a large amount of data. 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 (Salazar-Reyna et al., 2020). Healthcare industry searches for suitable technologies to streamline resources for the sake of improving the patient experience and organizational performance (Khanra et al., 2020).

On the other hand, data is considered as the "new oil" and the most important asset that currently drives or even dictates the future of science, technology, and the economy (Nielsen, 2017). Data science is defined from high-level perspective as the science of data or the study of data. The origins of data science can be traced back to the data analysis field in statistics and mathematics. Then, with the introduction of data mining and machine learning, data analysis became known as "data analytics". Data analytics is, in fact, a multidisciplinary field that examines and extracts new insights and conclusions from data. It is increasingly used for a variety of data and domain-specific tasks, such as business analytics, risk analytics, and web analytics, among others. Domain-specific analytics is at the heart of data science innovation and application in order to aid in knowledge discovery and decision-making (Salazar-Reyna et al., 2020). It provides means for extracting meaningful information and useful knowledge and insights from (massive amounts of) raw data, generally to guide (business) processes and to reach (organizational) goals or solve problems. In healthcare, data science has the potential, among others, to improve care, save lives and lower costs by identifying associations and understanding trends and patterns within the data. Healthcare data generally incorporate electronic medical records (EMRs) like patient's medical history, physician notes, clinical reports, biometric data, and other medical data related to health (Yadav et al., 2018). To grasp benefits from the power of such data, healthcare has being digitized. According to Gartner, "by 2024, healthcare providers that have adopted a digital health platforms approach will outpace competition and partners by 80% in the speed of digital transformation and new feature implementation" (Singh et al., 2021). Data scientists are putting their best effort to find valuable insight from the healthcare data for quality medical services (Raja et al., 2020). In this context, many research works were proposed applying data science to the various healthcare fields, including but not limited to pulmonology, cardiology, oncology and patient mortality. Several reviews tried to summarize them such as de la Torre Díez et al. (2016); Imran et al. (2020); Galetsi and Katsaliaki (2020a,b); Jiang et al. (2019); Khan et al. (2022); Khanra et al. (2020); Kruse et al. (2016); Miah et al. (2022); Pashazadeh and Navimipour (2018); Raja et al. (2020); Sabharwal and Miah (2022); Salazar-Reyna et al. (2020). However, either they do not fully conform the literature systematic review guidelines, suffering therefore from some flaws such the quality of the filtering process or the well-definition of a driving research question; or focus on a particular aspect of big data analytics in healthcare, such as the variety of data or the target problem. None of them systematically studied the literature to give a big picture, providing a founded means to identify research opportunities and challenges.

To overcome this gap, we conducted a systematic review, following Petersen et al. (2008, 2015) guidelines, and driven by the main research question: "which information contributes to understanding and getting a comprehensive overview of data science in healthcare domain?". The rest of this paper is structured as follows: Section 2 describes the research methodology and sets the research questions. Section 3 answers the research question. Section 4 reports the results analysis. Section 5 discusses the related work. Section 6 describes the threats to validity. Finally, Section 7 concludes the study and summarizes the opportunities and open issues.

2. Research questions and methodology

This review aims at presenting a comprehensive overview of the research topic, which is the use of data science within healthcare applications, driven by twofold objectives: (1) to construct a rich big picture of data science application practices in the healthcare field based on grounded evidence, and (2) to provide a founded means to identify research opportunities in area of interest. To achieve these objectives, we followed a systematic review process adapted from (Petersen et al., 2008, 2015) in order to collect existing works, analyze them to extract, summarize and synthesize information related to our defined research questions. These latter, as well as the details of the methodology, are presented in the following subsections.

2.1. Research questions

We defined a driving research question to reach the main objective as: ”which information contributes to understanding and getting a comprehensive overview of data science in healthcare domain”, from which we formulated three main ”W” research questions answering the ”What”, ”Why” and ”How” as follows:

1. RQ1: *What is the goal from using data science in healthcare domain?* In order to cover this research question, we derived three sub-questions dealing with the analytics goal and type as well as the field of application:
   - RQ1.1: What kind of data science approaches are applied in the healthcare domain?
   - RQ1.2: What type of insight is provided?
   - RQ1.3: In which healthcare fields data science is applied ?
2. RQ2: *Why data science is used in healthcare applications?* In order to cover this research question, we derived three sub-questions dealing with the targeted problems, the input and the output of such applications:
   - RQ2.1: What type of problems are targeted by data science in healthcare?
   - RQ2.2: What kinds of data variety and form are used?
   - RQ2.3: What types of data products are produced?
3. RQ3: How data science is applied in healthcare domain? In order to cover this research question, we derived three sub-questions dealing with used techniques, algorithms and targeted data science steps:

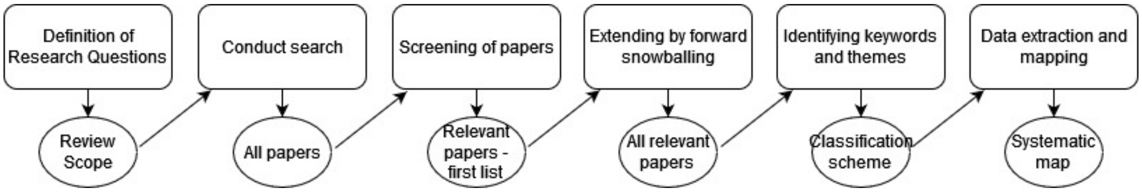
Definition of Research Questions
Review Scope
Conduct search
All papers
Screening of papers
Relevant papers - first list
Extending by forward snowballing
All relevant papers
Identifying keywords and themes
Classification scheme
Data extraction and mapping
Systematic map

Fig. 1. Review conducting process.

- RQ3.1: what kind of data science techniques are used?
- RQ3.2: what kind of machine learning algorithms are used?
- RQ3.3: Which data science steps are addressed?

2.2. Methodology

We carried out this systematic review using a well-defined process according to Petersen et al. (2008, 2015) guidelines, depicted in Fig.1. This latter shows that after defining the research questions, already presented in the previous subsection 2.1, it was question to conduct search in order to identify potential papers, that had later to be screened to keep only ones that are relevant to our study. This list was extended by a set of papers identified after applying a forward snowballing on the selected papers, to constitute the set of all relevant papers, or the set of primary studies. Afterwards, we defined a classification scheme based on a set of parameters (keywords and terms) in order to help in collecting information and therefore analyzing the primary studies to answer research questions. Obviously, the design of the classification schema is driven by the research questions. Main worth reporting details related to the search and study selection sub-processes are given in the following subsections.

2.2.1. Search sub-process

We advocated a primary studies searching and selecting strategy slightly different from what it is widely known in such review works. Indeed, since there are existing systematic reviews in the field of healthcare dealing with topics related to data science, it is wise to rely on them to identify the first list of papers as if it is returned from digital library sources. What also drove us to

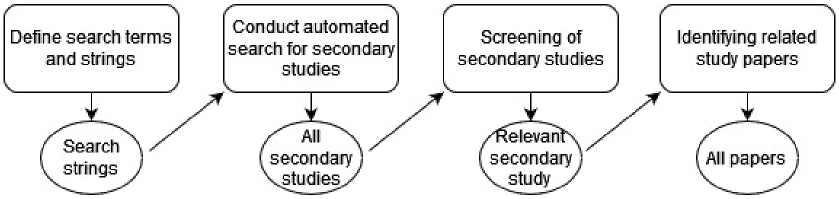
Define search terms and strings
Search strings
Conduct automated search for secondary studies
All secondary studies
Screening of secondary studies
Relevant secondary study
Identifying related study papers
All papers

Fig. 2. Search sub-process details.

consider such way of work is that the number of published research studies in the field of healthcare is huge and it would take a lot of time for us to screen returned papers from digital sources to keep relevant ones. Hence, the aim of search sub-process is to select a secondary study covering our focus area as a source of primary studies.

Fig.2 shows the tasks carried out to perform the search sub-process. First of all, it is question of defining search terms and strings. With respect to the focus of our review and the defined research questions, we advocated three key terms which are Kt1: ”healthcare”, Kt2: ”data science” and Kt3: ”systematic review”; each of them was extended by a set of synonyms picked from existing research works in the fields to obtain three sub-strings S1, S2 and S3, as illustrated by Table 1. The resulting search string is therefore a conjunction of these sub-strings in the form of *S1 AND S2 AND S3*.

The following step was to conduct an automated search for secondary studies using the defined search string. We selected a set of resources among most common and well known scientific digital libraries namely, ACM Digital Library, IEEE Xplore, Google Scholar, PubMed, Scopus and Science Direct. The search string was submitted to each of them and eventually adapted to the related search engine requirements. Fig.3 shows the number of papers returned by digital resources. The initial search resulted in a total of 151 papers, among them 18 were duplicated. The screening process started by applying a selection by title. 8 studies were kept from the 133 non duplicated ones. After reading the abstract and full-text, we identified only 4 secondary studies conforming the requirements, namely (Galetsi and Katsaliaki, 2020b; Khanra et al., 2020;

Table 1: Search sub-strings

No	Key term	Search sub-string	Source
1	healthcare	healthcare OR health OR medicine OR medical OR hospital OR clinical OR clinic OR disease	(Sadoughi et al., 2020)
2	data science	"data science" OR "data analytics" OR "data mining" OR "data processing" OR "big data" OR "big data analytics"	(Fernández Del Carpio and Angarita, 2018)
3	systematic review	"systematic review" OR "literature review" OR "systematic mapping" OR "meta review" OR "literature analysis" OR "literature survey" OR "meta analysis" OR "structured review"	(Kitchenham et al., 2010)

Raja et al., 2020; Salazar-Reyna et al., 2020). We therefore proceeded to a selection by quality venue, using the quartile, the SJR and IF as indicators. The publication year was not considered as all of them were published in 2020. Finally, (Salazar-Reyna et al., 2020) was selected as secondary study, from which we got the first list of research studies for our review.

2.2.2. Study selection sub-process

The followed process to select relevant primary studies is described in Fig. 4. 471 papers were retrieved from the identified secondary study. Firstly, we extracted only journal venues; which means that we excluded all papers that are published in conferences or workshops or as a book or a book chapter. Then, we removed duplicated venues. A set of 208 journal venues underwent a quality assessment. Indeed, we defined three exclusion criteria as: E1: non ISI, E2: emerging and E3: Q3 or Q4 quartiles. The resulted list of 131 journals were filtered according to the topic that should be related to computer science and/or healthcare domains. Hence, only 76 high quality Q1 and Q2 quartiles journals

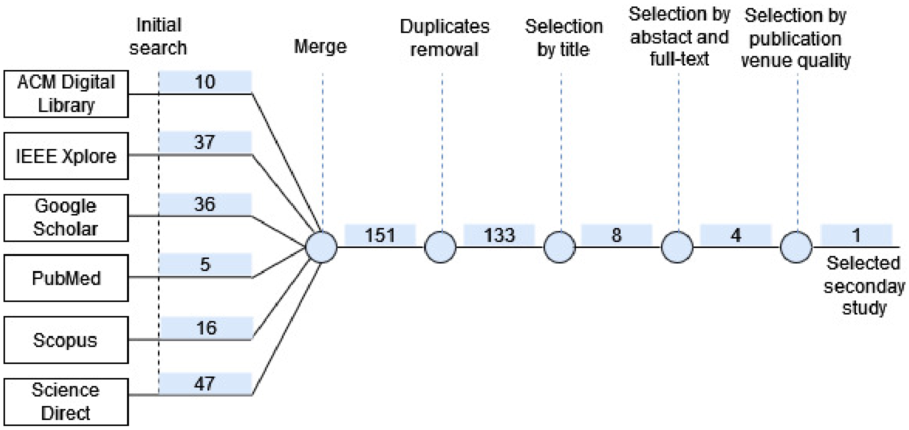

Fig. 3. Secondary study selection details.

were retained[1]. This latter journals list were used to screen the 471 papers, resulting in 153 research works. In the next step, we used the titles and read the abstracts to keep 87 potential papers. After full-text reading, 62 relevant primary studies were identified.

Since, the secondary study does not cover the whole 2019 year, neither 2020 nor 2021 years, we used the forward snowballing in order to extend the relevant papers list. As a matter or fact, for each of the 62 primary studies, we used Google scholar to determine how many times it was cited and retrieve necessary information about the citing papers. We identified 2354 citing papers that underwent the same process than the first ones. 149 new relevant primary studies published between 2019 and 2021 were identified thanks to this process. Therefore, a total of 211 relevant primary studies were identified at the end of the study selection process[2]. We thoroughly read and analyzed each of them to extract information according to the defined classification scheme, in order to answer the research questions. Results are presented in the following section.

[1] The list of retained journals is available on https://cutt.ly/E8SCDNu #sheet1

[2] The full list of retained primary studies is available on https://cutt.ly/E8SCDNu #sheet2

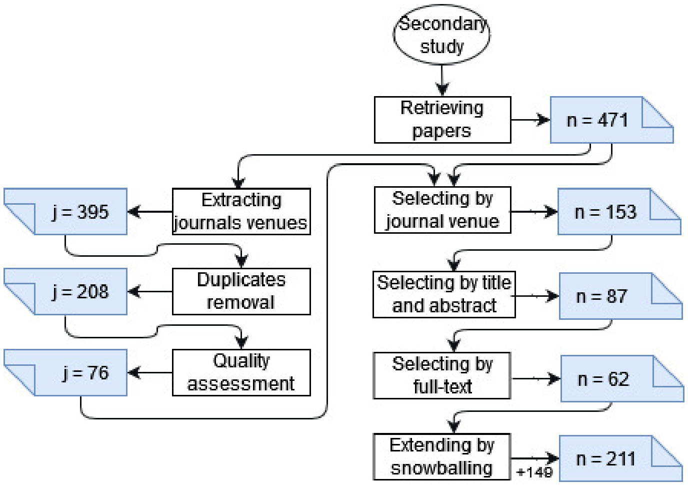

Fig. 4. Primary study selection details.

3. Results

3.1. RQ1. What is the goal from using data science in healthcare domain?

3.1.1. RQ1.1. What kind of data science approaches are applied in the healthcare domain?

According to Cao (2017a), there are in general four kinds of data science approaches tightly related to the goal to achieve.

- *Understanding*: when analytics are essentially based on historical data to explore ”what happened” and to gain insights into ”how and why it happened”. Methods, such as modeling and experimental design, allow to reach the fundamental goal, which is to undertake a reactive understanding of what took place.

- *Detection*: when analytics are essentially based on present data to explore ”what is happening” in order to derive insights about ”how and why it happens”. Methods, such as alerting about suspicious events or detecting interesting groupings or patterns, may be used to reach the fundamental goal, which is extracting insights for decision making purposes.

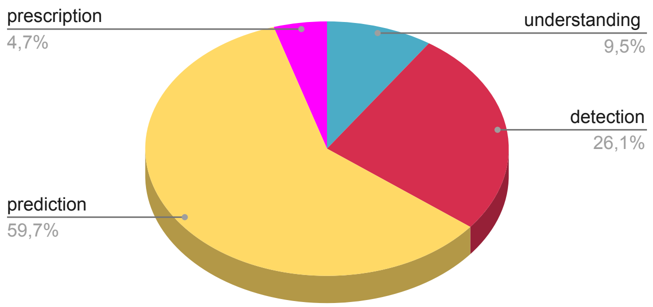

Fig. 5. Kind of data science applied approaches.

- *Prediction*: when analytics are essentially based on future data to investigate "what will happen" and to generate insights into "how and why it will happen". Methods, such as estimating the occurrence of future events, grouping, and patterns, allow to reach the fundamental goal, which is to achieve a proactive understanding, forecasting, and prediction, as well as early prevention.

- *Prescription*: when analytics are based on past, present and future data to investigate "what best action to take" and interpret findings in order determine the optimal actions to undertake. Making optimal recommendations and actionable interventions allow to achieve the fundamental goal, which is actively and optimally solve the identified problems through actionable decisions.

Fig.5 shows that "prediction" is the most kind of data science applied approaches in the healthcare domain (59.7%), followed by "detection" (26.1%); while only 9.5% of works applied an "understanding" approach and 4.7% applied a "prescription" approach.

3.1.2. RQ1.2. What type of insight is provided?

According to Cambridge dictionary, an insight is defined as "(the ability to have) a clear, deep, and sometimes sudden understanding of a complicated problem or situation"[3]. As to Oxford dictionary, it stated that an insight is "the ability to see and understand the truth about people or situations" and explains an "insight (into something) as an understanding of what something is like"[4]. In the context of data science, the word insight refers to the value acquired through the application of analytics. In fact, it is about the discovery of a pattern or a relationship between variables not previously known, gained by analyzing data and information in order to understand the context of a particular situation, leading to drawing conclusions, making actions and taking actionable decisions. The insight is characterized as (Cao, 2017a):

- *reactive*: when it is gained by analyzing previous situations in order to understand what happened.
- *active*: when it is gained to respond to currently happening events or to issues as they occur.
- *proactive*: when it is gained in order to prevent problems from occurring.
- *on-demand*: when it is gained to enable optimal actions and recommendations in order to plan actionable decisions.

Fig. 6 shows that more than the half of works (58.3%) using data science in healthcare domain aim at avoiding problems before they arise by providing a proactive insight. Almost the quarter of them (22.7%) address occurring problems as they provide active insight. The tenth of works (10.4%) aim at gaining better understanding of what happened by providing reactive insight. Only 3.3% of studies focus on on-demand insight, namely Baytas et al. (2016); Choi et al. (2016); Guo et al. (2016); Hu et al. (2020); Waqar et al. (2019);

[3] https://dictionary.cambridge.org/dictionary/english/insight

[4] https://www.oxfordlearnersdictionaries.com/definition/english/insight

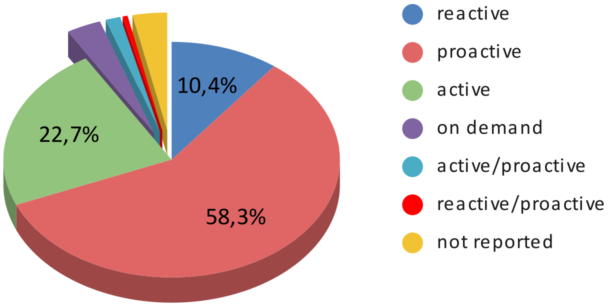

Fig. 6. Provided insight types in healthcare domain.

Yang et al. (2020); Ye et al. (2019). Very few works used analytics to provide two types of insights i.e. reactive and proactive (Eyuboglu et al., 2021), and active and proactive (Alharithi et al., 2021; Ning et al., 2021; Wong et al., 2017). Finally, it is worth mentioning that 7 works (3.3%) have not reported any type of insight.

3.1.3. RQ1.3. In which healthcare fields data science is applied?

Several fields are targeted by the use of data science in healthcare domain. We identified 26 fields within the 211 works subject of this study. As shows Fig. 7, cardiology (e.g. Nasir et al. (2018); Nicholson et al. (2021)) and healthcare services (e.g. Shah et al. (2020); Waqar et al. (2019)) are on top of the list with 29 works each, representing 12.8%, followed by the oncology field with 22 works (9.7%) (e.g. Ganggayah et al. (2019); Nayan et al. (2022)), then pulmonology with 17 works (7.5%) (e.g. Khasha et al. (2021); Xiang et al. (2020)). Next, come the neurology (e.g. Raghavaiah and Varadarajan (2021); Razavi et al. (2019)), diabetology (e.g. Alfian et al. (2020); Massaro et al. (2019)), patient mortality (e.g. Beeksma et al. (2019); Blom et al. (2019)), hospital readmission (e.g. Ashfaq et al. (2019); Baechle and Agarwal (2017)), ophthalmology (e.g. Cao et al. (2020); Pendleton et al. (2021)) and palliative care (e.g. Avati et al. (2018); Guo et al. (2021)) fields with a percentage between 6.6% representing 15 works and 4.4% representing 10 works. The fields that received the least

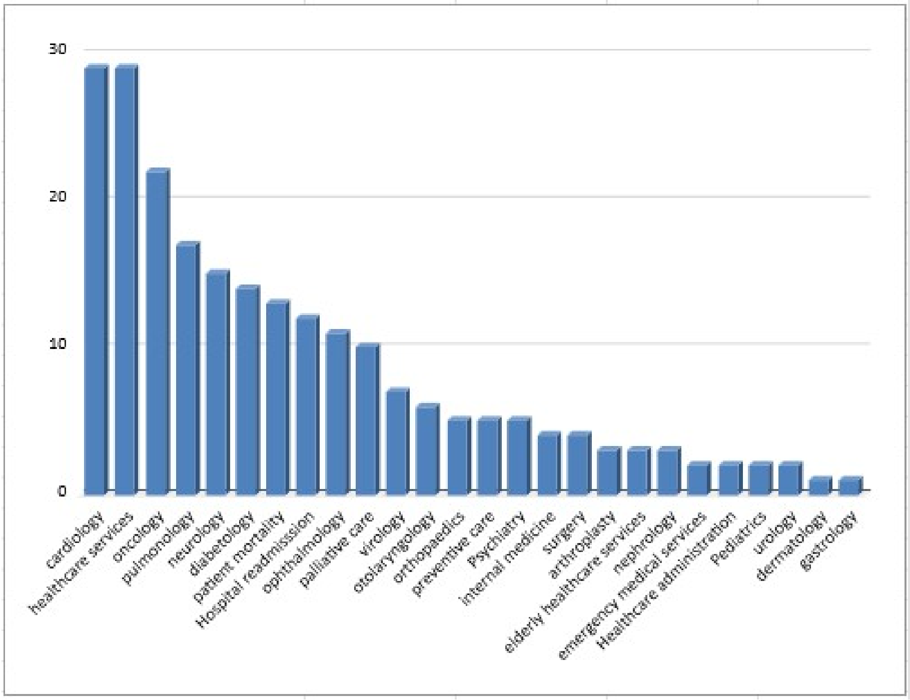

Fig. 7. Data science targeted healthcare fields.

interest are emergency medical services (i.e. Chao-Wen et al. (2015); Sadrawi et al. (2018)), healthcare administration (i.e. Crump et al. (2021); Liu et al. (2016)), pediatrics (i.e. Gálvez et al. (2017); Li et al. (2019b)), urology (i.e. Liu et al. (2015, 2021)) with 2 works each (0.9%), while the dermatology (i.e. Srinivasu et al. (2021)) and gastrology (i.e. Nudel et al. (2021)) fields were the focus of only 1 work each. It is worth highlighting that almost 7.5% (16) of the studied works targeted two fields, especially diabetology and ophtalmology given the relationship between the two diseases.

3.2. RQ2. Why data science is used in healthcare applications?

3.2.1. RQ2.1. What type of problems are targeted by data science in healthcare?

Several types of problems may be targeted by data science as stated by Dorr et al. (2016). Fig. 8 shows that almost half of the works (48.2%) aim at *prediction* by estimating the future of some variable of interest, such as heart failure in

Lu et al. (2021); Wang et al. (2020), hypertension in Lopez Bernal et al. (2021a); Chang et al. (2021); Kanegae et al. (2020); Ye et al. (2018), and 30-days patient mortality in Blom et al. (2019); MacKay et al. (2021). 29% of the works are related to *classification* as they focus on determining whether a patient has some disease such as breast cancer in Alhamid (2019); Chen et al. (2019); Ganggayah et al. (2019), or brain related diseases like alzheimer in Razavi et al. (2019), brain tumour in Jayalakshmi and Rao (2020), and brain disorder in Kale et al. (2019). Furthermore, some of these works consider classifying patients in order to avoid hospital readmission which is the case of Baechle and Agarwal (2017); Jain et al. (2019). Another type of problems targeted by data science in healthcare is *detection*, where the datasets are analyzed in order to find data of interest. 6.3% of the works fall in this context, almost half of them used detection as part of doctor or physician recommendation (Guo et al., 2016; Hu et al., 2020; Waqar et al., 2019; Yang et al., 2020; Ye et al., 2019; Yuan and Deng, 2021). Similarly to detection, *knowledge base construction* in specific healthcare field and specific disease is dealt with by 14 works. Different fields are targeted by these works, such as oncology (cancer) in Afzal et al. (2017); Nicholson et al. (2021), pulmonology (asthma) in Khasha et al. (2019); Toti et al. (2016) and internal medicine (chronic inflammatory diseases) in Veroneze et al. (2020). As to *data fusion* which aims at integrating different representations, it aroused the interest of researchers in only 3.6% of the works. 2 of them are exclusively dedicated to this problem class i.e. (Pustišek, 2017; Wimmer et al., 2016), whereas the 6 others consider one or more other problem types in addition to data fusion. We cite as example prediction in Han et al. (2017); Shen et al. (2021), classification in Khasha et al. (2021), and classification and knowledge base construction in Tsai et al. (2016). In 3.1% of the works, researchers focused in *detecting anomalies* in datasets related to neurology (Shanker and Bhattacharya, 2021; Zhang et al., 2016), to otolaryngology (Harar et al., 2020; Wang et al., 2019) and healthcare administration (Liu et al., 2016). 2 works, namely Chang (2018) and Wong et al. (2017) respectively tackled data fusion and prediction problems in addition to anomaly detection. 4 works focused in

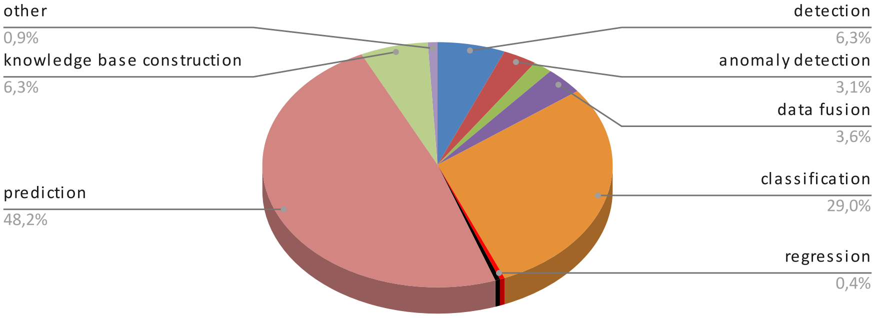

Fig. 8. Problems types targeted by data science in healthcare.

eliminating errors and inconsistencies in data, which is known as *cleaning*. 3 among them are proposed within the healthcare services field (Abdullah et al., 2020; Baytas et al., 2016; Woo et al., 2019). Finally, *regression* and *alignment* are the object of 1 work each, respectively Zhang et al. (2019) and Yang et al. (2021).

3.2.2. RQ2.2. What kinds of data variety and form are used?

As mentioned earlier, insights are gained thanks to analytics. Nevertheless, analytics are inconsequential without data. Indeed, data is the core of data science. It may be collected from various and heterogeneous data-sources such as medical imagery, IoT devices and social networks. It may therefore take one of the known forms: *text*, *image*, *audio* or *video*. When used in its raw format, that's to say audio, image, video and text-heavy files, data is referred to as *unstructured*. In contrast, when it is organized following a given model and according to a specific format, such as relational data and numerical information, data is referred to as *structured*. When data is not conform to relational representation, but still has some structure like hierarchy of records based on key-value tags, it is referred to as *semi-structured* (Group).

The analysis of the studied works revealed that 25.4% of them used unstructured data, 63.4% were based on semi-structured data, while 11.2% used structured data. We discuss in the sequel this distribution with respect to data forms illustrated in Fig. 9. As a matter of fact, we identified 5 works (about 0.5%) that used audio data as input, all of them focusing on voice pathology within the otolaryngology healthcare field (Alhussein and Muhammad, 2019; Harar et al., 2020; Hossain and Muhammad, 2016; Kadiri and Alku, 2019; Verde et al., 2019). For the experimentation purposes, these works used various known voice databases, namely Saarbruecken Voice Disorders Database (SVD) for all works, Massachusetts Eye and Ear Infirmary (MEEI) database for Alhussein and Muhammad (2019); Harar et al. (2020); Hossain and Muhammad (2016); Verde et al. (2019); the Hospital Universitario Prıncipe de Asturias (HUPA) database for Harar et al. (2020); Kadiri and Alku (2019); in addition to VOice ICar fEDerico II (VOICED) database for Verde et al. (2019); and Arabic Voice Pathology Database (AVPD) for Harar et al. (2020). 4 among the 5 works extracted various features like IDP, MFCC, PLP, jitter, shimmer and HNR, from the signal to detect voice pathology (Harar et al., 2020; Hossain and Muhammad, 2016; Kadiri and Alku, 2019; Verde et al., 2019); while Alhussein and Muhammad (2019) transformed the voice into spectro-temporal representation.

As to video data as input, it was used in a single work in which echocardiographic videos were used to improve predictions of all-cause mortality (Ulloa Cerna et al., 2021). For this purpose, the authors used two clinical databases of echocardiographic videos: Philips iSite and Xcelera.

Almost 10% of the studied works were based on various images types as input. For example, in the neurology healthcare field, magnetic resonance brain images (MRI) datasets were used, especially ADNI in Raghavaiah and Varadarajan (2021); Razavi et al. (2019), PBDS in Zhang et al. (2016), and DS-nn from Harvard Medical School in Jayalakshmi and Rao (2020); Kale et al. (2019); Shanker and Bhattacharya (2021); while in the virology field, we find Chest computer tomography (CT) imaging used in Castiglione et al. (2021) through the SARS-COV-2 CT-Scan dataset from Kaggle and a Chest Radiography Images (X-ray) used in Alharithi et al. (2021); Elakkiya et al. (2021). As to functional magnetic resonance images (fMRI), they were used in Tahmassebi et al.

(2018) and Nozais et al. (2021). In Sadrawi et al. (2021), the dataset image represents electrocardiography (ECG) signals in the context of cardiovascular and cerebral hemodynamics. Raw retinal color fundus images from the wo popular datasets MESSIDOR and IDRiD in the ophtalmology healthcare field were used in Saranya and Prabakaran (2020).

While data variety in all previous forms is unstructured as works deal with audio, video and images, textual form of data, which is the most used (86.6%), may take any of the three variety kinds as shown in Fig. 9:

- structured in 25 works; for example, in the oncology healthcare field, Richter and Khoshgoftaar (2019a) used data collected from a mobile-first structured data-input, and cloud-based dermatology-specific electronic health record system to predict melanoma risk (Richter and Khoshgoftaar, 2019a,b). Krempel et al. (2018) developed a database of cancer-related data that can be used for predictive cancer classification. As to Nicholson et al. (2021), they proposed an ontology to model the international rules for multiple primary cancers in description logic based on a common European data set conform to the European Network of Cancer Registries. In other fields such as surgery, Zhang et al. (2018) based their research on data coming from Vanderbilt's Perioperative Data Warehouse in the context of perioperative patient acuity, while in neurology Djatna et al. (2018) utilized data from BioMed Central to deal with stroke disease.

- semi-structured in 142 works; information is coded as attribute-value; it is related to (i) demographic data such as age, gender, height, smoking status, as well as (ii) medical data about patient history through some given variables or examination indicators; this includes also treatment history according to the diagnosis. In the context of cardiology field, which was the subject of 21 studies, variables may be for example blood pressure, heart failure and blood biochemistry. We identified three different sources of such data: (1) electronic medical records (EMR) composed of data collected in one healthcare organization i.e. a hospital, a clinic, or clinician's in one practice which is cardiology (Alshakhs et al., 2020; Chang et al., 2019, 2021; Hyland et al., 2020; Kanegae et al., 2020; Lu and Uddin, 2021; Narayan and Sathiyamoorthy, 2019; Nasir et al., 2018; Ross et al., 2016; Wang et al., 2020; Wu et al., 2020); (2) electronic health records (EHR) containing information from several healthcare providers involved in the patient's general care such as laboratories and nursing home in addition to hospitals or clinics. Based on these records, desired attributes related to cadiology and estimated valuable according to the aim of the study are extracted to form the input dataset (Cho et al., 2019; Datta et al., 2020; Han et al., 2017; Lu et al., 2021; Nudel et al., 2021; Ross et al., 2019); and (3) sensors including wearable devices and biosensors and/or mobile Web applications where data is personally introduced (Lopez Bernal et al., 2021b; Pustišek, 2017; Yoo et al., 2020; Zhang et al., 2019).

- unstructured in 27 works; in healthcare services field for example, Zhao et al. (2018) used textual notes of clinical images such as body system, image modalities, clinical findings and case discussion in order to generate a knowledge model that represents the contents of medical imaging reports. As to Li et al. (2019a), they proposed to extract valuable knowledge from the clinical free-text stored in the Electronic Medical Records. Hao and Zhang (2016), Yang et al. (2016) and Hu et al. (2020) based their works on textual data like reviews about physicians and medical stories collected from Web specialized communities and forums, resp. Good Doctor Online platform (Hao and Zhang, 2016), MedHelp (Yang et al., 2016) and Haodf.com (Hu et al., 2020) in order to target patients concerns and/or stakeholders concerns. In the same context, Fairie et al. (2021) used a dataset from FACT database where telephone calls to the health authority were summarized in free-text narrative style by patient concerns consultants. In order to recommend the right doctor, Guo et al. (2016) used more various data, such as papers in scientific journals, presentation activities, besides patient reviews from multiple Web sources, to construct doctors footprints. Finally, Hu et al. (2021) used clinical notes, especially free-text discharge summaries from the MIMIC III database to predict medical codes.

Seven (7) works among the preceding studies used more than one data variety. Indeed, 5 studies used semi-structured as well as unstructured data. The semi-structured data comes from electronic medical records and eventually the International Classification of Diseases (ICD) coding system in (Beeksma et al., 2019; Murphy et al., 2019; Weegar and Sundström, 2020; Yang et al., 2020) and from collected profiles of physicians specialized in hypertension in (Ye et al., 2019), while the unstructured data comes either from the free-text based clinical notes or reviews. As to Golas et al. (2018), they used structured as well as unstructured data extracted from two data warehouses to predict the risk of 30-day readmissions in patients with heart failure. Finally, Makino et al. (2019) predicted the progression of diabetic kidney using unstructured, semi-structured and structured data including lab tests, medication, anamnesis, diagnosis treatment and longitudinal data.

It is worth highlighting that 4 works were based on input comprising more than one kind of data: (1) 3 among them used text and image to achieve their goal, namely (i) Periasamy et al. (2022) where data consists of semi-structured patient-related information like age, calcium level, smoking or drinking habit and the previous medical treatment history as well as patients image data composed of computed tomography (CT) scans, both serving to predict osteoporosis; (ii) Shah et al. (2020) analyzed the patients' sentiment regarding the quality of healthcare service delivery using the unstructured text and photographic information in the online reviews from the Yelp.com platform; (iii) Eyuboglu et al. (2021) proposed to detect abnormality in whole-body based on unstructured radiology reports and F-fluorodeoxyglucose (FDG) positron emission tomography (PET)/CT scans; and (2) 1 work proposed an intelligent navigation technique for accurate image-guided Covid-19 lung biopsy based on chirurgical audio, medical images (CT scans) for Covid-19 of lungs and semi-structured biomechanical

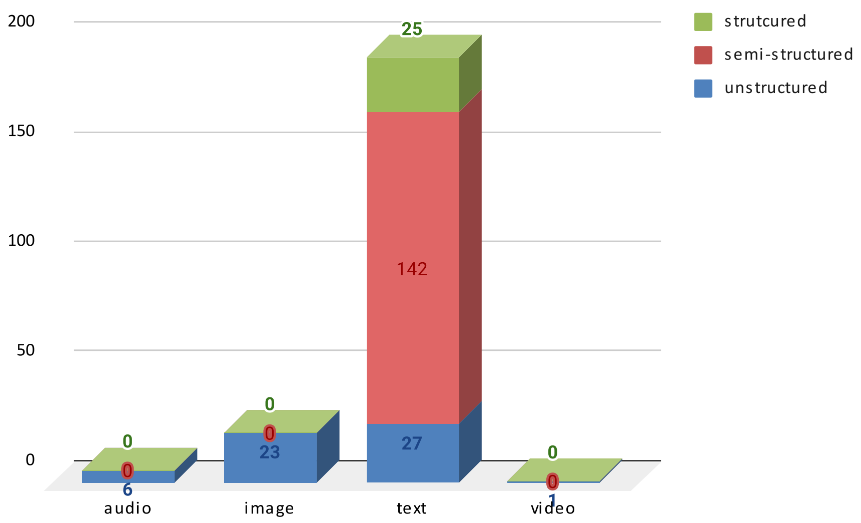

Fig. 9. Input data variety and forms used by data science in healthcare.

textual data (Tai et al., 2021).

3.2.3. RQ2.3. What types of data products are produced?

D.J. Patil, a chief data scientist, defines a data product as ”a product that facilitates an end goal through the use of data” (Patil, 2012). Data is transformed through a process involving ’think with wisdom’, ’understand domain’, ’manage data’, ’compute with data’, ’mine on knowledge’, ’communicate with stakeholders’, ’deliver products’, and ’act on insights’. (Cao, 2016). Rowley (2007) sees therefore that the aim is to deliver *knowledge*, *intelligence*, and *wisdom*; Cao (2017a) adds *decision* and highlights that these are the ultimate values of data products. Knowledge represents the form of processed information in terms of an information mixture, procedural actions, or propositional rules (Cao, 2017b); it is often construed as know-how (besides the know-that) (Frické, 2018). Wisdom is defined as ”the ability to use knowledge and experience to make good decisions and judgments”[5]. Actually, it requires to use knowledge

[5] https://dictionary.cambridge.org/dictionary/english/wisdom

and insights gained from the information in order to prevent problems from occurring through answering questions such as 'why do something' and 'what is best' (Ontotext); it is seen as how to use the knowledge into practice (Jifa, 2013). As to intelligence, it is defined as "the ability to learn, understand, and make judgments or have opinions that are based on reason"[6]. It is the product of analyzed and synthesized information creating insights (Rabkin, 2016). Finally, decision refers to a choice made about something after thinking about several possibilities[7].

Fig. 10 shows that 82.5% of studied works aimed at producing knowledge. All healthcare fields were concerned with such output type. Only 7.6% (16 works) targeted intelligence type data product in various healthcare fields such as patient mortality (Bergquist et al., 2020; Choi and Boo, 2020; Sahni et al., 2020), pulmonology (Arefeen et al., 2020; Toti et al., 2016; Xiang et al., 2020), and healthcare services (Abdullah et al., 2020; Almasoud et al., 2019; Zhao et al., 2018); while 5.7% (12 works) targeted decision type data product, forth of them are within the oncology healthcare field (Afzal et al., 2017; Alhamid, 2019; Richter and Khoshgoftaar, 2019a; Wimmer et al., 2016). Wisdom received the least focus with 9 works (4.3%), most of them are within healthcare services field, especially doctor recommendation (Guo et al., 2016; Hu et al., 2020; Waqar et al., 2019; Yang et al., 2020; Ye et al., 2019; Yuan and Deng, 2021); and 1 work related to cardiology (Hyland et al., 2020).

3.3. RQ3. How data science is applied in healthcare domain?

3.3.1. RQ3.1. what kind of data science techniques are used?

In addition to foundational knowledge in the applied field required to be able to harness value from data, data science combines aspects of mathematics, statistics, visualization and programming to automatically analyze enormous amount of data and extract information with the aim of making new discoveries

[6] https://dictionary.cambridge.org/dictionary/english/intelligence

[7] https://dictionary.cambridge.org/dictionary/english/decision

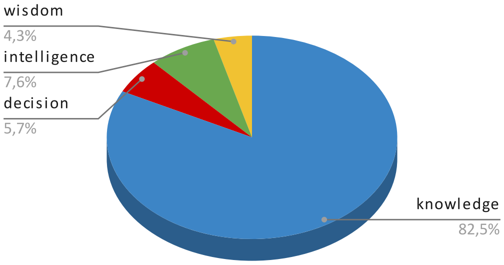

Fig. 10. Data products types produced by data science in healthcare.

(Probyto, 2020). To that end, one or several techniques may be used such as regression, time series analysis, neural networks and multivariate analysis relying on star plots. These techniques are either based on *mathematics and statistics*, or based on *artificial intelligence* including *machine learning*, or else based on *visualization and graphs*. The study of the 211 papers revealed that almost 93% used a single technique while the rest combined two techniques from the previously mentioned ones (c.f. Fig. 11). In fact, 87.2% of the works used artificial intelligence and machine learning; they covered all identified healthcare fields (cf. section 3.1.3). Various machine learning algorithms have been used in these studies; we cite as examples Naive Bayes in Alharthi et al. (2019); Jain et al. (2019); Liu et al. (2015); XGBoost in Chang et al. (2021); Li et al. (2021); Nayan et al. (2021); Support Vector Machines (SVM) in Alhamid (2019); Kadiri and Alku (2019); Polce et al. (2021); Random Forest in Cho et al. (2020, 2021); Sahni et al. (2020); Long Short-Term Memory (LSTM) in Beeksma et al. (2019); Shen et al. (2021); Ye et al. (2019); Logistic Regression in Datta et al. (2020); Fritzell et al. (2021); Makino et al. (2019); Ensemble Learning in Elakkiya et al. (2021); Khasha et al. (2019, 2021); Wang et al. (2020); and Neural Networks with their variants: Multi-Layer Perceptron in Nozais et al. (2021); Zhang et al. (2016), Recurrent Neural Networks in Lu et al. (2021); Tomašev et al. (2021), Convolutional Neural Networks in Periasamy et al. (2022); Ulloa Cerna et al. (2021), Deep Neural Networks in Chen and Wu (2020); Raghavaiah and Varadarajan

(2021). Some studies used more than one machine learning algorithm among the above mentioned ones in their works, such as Batool et al. (2016); Gálvez et al. (2017); Goto et al. (2019); Weegar and Sundström (2020); Wimmer et al. (2016). Others, added to the combination other algorithms like Decision Trees in Ganggayah et al. (2019); Sena et al. (2019); Zhang et al. (2021), and K-Nearest Neighbors in Blom et al. (2019); Cao et al. (2020); Wong et al. (2019). On the other hand, 7 works representing 3.3% used mathematics and statistics techniques; among them in the healthcare field, Fairie et al. (2021); Hao and Zhang (2016); Yang et al. (2016) relied on the statistical model Latent Dirichlet Allocation (LDA) for topic modeling, and Almasoud et al. (2019); Feldman et al. (2015) were based on semantic similarity calculations. Furthermore, 2.7% representing 5 works used visualization and graphs techniques. Pustišek (2017) in the cardiology field provided contextualized e-health dashboards and alerting; Baytas et al. (2016) in healthcare services proposed visual analytic tool to interactively build and navigate a phenotype hierarchy; Chang (2018) in the oncology field targeted simulating medical imaging as alternative to real medical imaging; Liu et al. (2016) in healthcare administration field developed graph-analysis techniques and applied them to look for fraud, waste, and abuse activities; and Pendleton et al. (2021) in ophthalmology field developed a disease specific ontology with its graphical visualization, enriched with patient-preferred synonyms. Regarding studies combining two techniques, we have identified that: (i) 4.74% (10 works) used artificial intelligence and machine learning with mathematics and statistics techniques; for example in the cardiology field, Fast Fourier Transformation was used to preprocess time series patient data in order to extract meaningful features serving as input of one or several machine learning algorithms, including neural networks, either to provide personalized heart condition classification (Yoo et al., 2020) or to predict chronic heart diseases (Narayan and Sathiyamoorthy, 2019); (ii) 2.7% (5 works) used artificial intelligence and machine learning with visualization and graphs techniques; for example, Zhao et al. (2018) in healthcare services field created a knowledge model by combining natural language processing and semantic net-

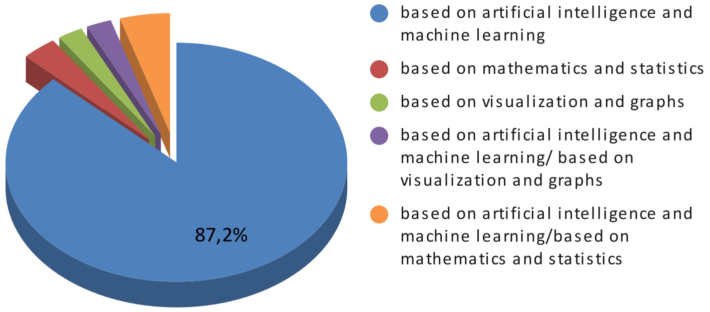

Fig. 11. Kind of data science techniques in healthcare.

work analysis; Lu and Uddin (2021) predicted chronic cardiology/pulmonology diseases based on Graph Neural Networks, while Foufi et al. (2019) in internal medicine field used text mining to extract entities and their relations, and built a graph of chronic diseases in order to map the chronic disease entities.

3.3.2. RQ3.2. what kind of machine learning algorithms are used?

As mentioned in the previous sub-section, several machine learning algorithms were used in the works using data science within healthcare domain; precisely 90% of them were built on one or more machine learning algorithms. Obviously, this type of works needs to rely on a dataset to make the algorithm learn from actual data in order to predict, to classify, to associate and so on. As known, three categories of learning approaches do exist:

1. *supervised* learning when the algorithm needs to be taught the output; it constructs a model from labeled training data, yielding to desired outputs from inputs. This type of learning, where algorithms such as Linear Regression, Naive Bayes, Support Vector Machine, Random Forest and Neural Networks may be used, is tightly closed to classification and regression problems.
2. *unsupervised* learning is used when the goal is to learn more about data. In other words, algorithms of this category, such as deep learning Convo-

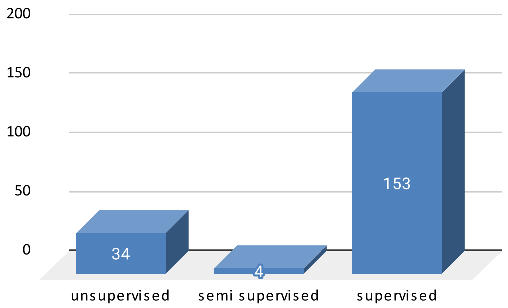

Fig. 12. Kind of machine learning algorithms used by data science in healthcare.

lutional Neural Networks Autoencoders, the Recurrent Neural Networks variant Long Short-Term Memory networks and K-means, are left to their own to discover some structure or distribution of data i.e. patterns, based on unlabeled data to solve association or clustering.

3. *semi-supervised* learning sits in between supervised and unsupervised learning approaches. It is used when just a small portion of data is or can be labeled as having recourse to human experts to label a huge amount of data is very expensive and time consuming. A mixture of supervised and unsupervised algorithms may therefore be used.

Fig. 12 shows that (1) 80.1% of the works relying on machine learning algorithms have used supervised learning; (2) 17.8% (34 studies) have used unsupervised learning, mainly implementing various types of deep neural networks such as Alakus and Turkoglu (2020); Cho et al. (2019); Sadrawi et al. (2021); Xiang et al. (2020); while only 2.1 % representing 4 studies, i.e. Ashfaq et al. (2019); Foufi et al. (2019); Huang et al. (2016); Murphy et al. (2019), have used semi-supervised learning.

3.3.3. RQ3.3. Which data science steps are addressed?

To turn a problem to a solution, data scientists follow a well-defined process helping in discovering unseen patterns, converting information to actionable insights and making decisions. This process involves six steps (Cielen and

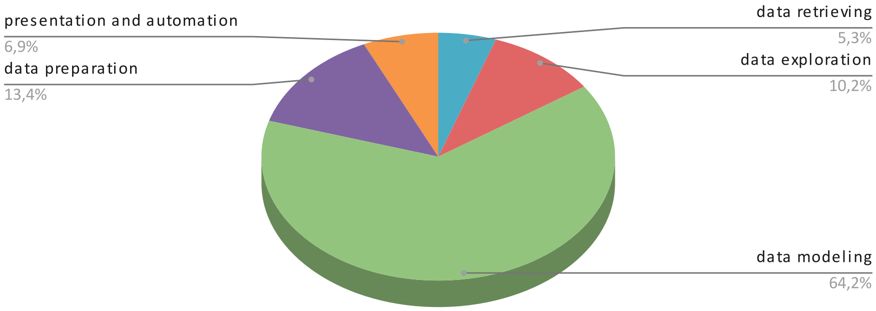

Fig. 13. Data science steps in healthcare.

Meysman, 2016): (1) *defining the problem* where the research goal is clearly set and the context is framed; (2) *data retrieving* where raw data is collected from various sources; (3) *data preparation* where raw data is checked and processed to eliminate errors like corrupted or missing values in order to obtain clean data; (4) *data exploration* where data is analyzed to develop ideas that can help discovering hidden patterns and insights; (5) *data modeling* where data is thoroughly and in-depth analyzed using mathematical, statistical and machine learning tools to build a model allowing to encompass every insight; and (6) *presentation and automation* where the results and insights are communicated through visualization techniques among others.

Fig. 13 shows that data modeling is the most targeted step with 64.2% of the works (e.g. Ganggayah et al. (2019); Nayan et al. (2021); Weegar and Sundström (2020)), followed by data preparation with 13.4% (e.g. Hu et al. (2021); Sadrawi et al. (2021); Wimmer et al. (2016)), then data exploration with 10.2% (e.g. Avati et al. (2018); Hao and Zhang (2016); Fairie et al. (2021)), presentation and automation with 6.9% (e.g. Chang (2018); Connolly et al. (2017))and finally data retrieving with 5.3% (e.g. Nasir et al. (2018); Tsai et al. (2016); Wong et al. (2017)). It is worth mentioning that data retrieving was not dealt with as a single targeted step, but with other steps including data modeling.

4. Discussion

The objective of this work was to assess and synthesize the published literature related to the application of data science in healthcare domain and to identify research opportunities. Based on the conducted study, we identified five key points described in the sequel.

First, the number of publications indicates that more than the half of identified works (59.7%) use data science for prediction in healthcare domain, however less than 5% of them focused on prescription. Prescription as analytics goal is not well addressed by authors and needs to be more investigated to make optimal recommendations and actionable interventions. It will be more interesting if data science studies in healthcare sectors not only predict what is most likely to happen in the future, but also recommend actions that practitioners can take to affect those outcomes.

Second, healthcare organizations rely on the power of data to gain valuable insight into patient health and treatment. According to our analysis, about 75% of the studied papers deal with structured and semi-structured data. Structured data is easily arranged in rows and columns and can be queried to return relevant search information. Semi-structured data is more complicated than structured data and cannot be stored in traditional databases. Unstructured health data remains one of the most precious and untapped resources in the healthcare ecosystem. Indeed, about 25% of published papers address unstructured data that is too complex to be parsed and interpreted. This data is any data type that cannot be stored in a relational database, including data in the form of images, videos, audio, Web pages, free text, and even social media content.

Third, an important finding is that 82.5% published studies targeted knowledge as data product. They were interested in adding meaning and value to the collected data. However, only 4.3% focused on wisdom, where the proposed healthcare solutions use the knowledge and insights gained from the information to take proactive decisions. Therefore, another challenging issue is transforming healthcare data into wisdom. Certainly, machine-learning algorithms try to imitate human intelligence and claim to understand the data. However, these algorithms are data centric and the phenomenon of human wisdom is absent. Our thorough literature review makes clear that researchers have long discussed artificial intelligence, but the problem of human wisdom in machine learning has not received much attention from the artificial intelligence research communities. This result is of value to both academics and practitioners illustrating that it is important to integrate artificial wisdom (AW) that enables a system to understand the whole environment to become adoptable for a wise decision.

Fourth, another interesting finding is that only 13.4% of the studies focused on the data preparation step of the data science process. Healthcare data can come in a variety of forms and contain errors, noise, missing numbers, and other irregularities. Such imprecise data could compromise the effectiveness of the study and produce false conclusions. Therefore, before using any prediction model or data analytics technique, it needs to be preprocessed, or structured and cleansed. Data preparation is a very time-consuming but crucial phase to overcome problems.

Fifth, artificial intelligence and machine learning based technique is the most used technique to apply data science in healthcare field. According to our review, more than 85% of the studies used such type of techniques. One of the drawbacks of the most of machine learning models is that results are presented as black box decisions. Recently, artificial intelligence researchers have concentrated on the creation of interpretable and explainable models as part of the explainable AI research subject. This type of model can also be advantageous in the context of healthcare applications. It is also obviously clear from our findings that supervised learning significantly overweighs unsupervised learning. This is due to the fact that more classification tasks are applied in healthcare field. It also shows that data is abundant in healthcare sectors, and as such, supervised learning can be used to provide more accurate results. Classification is the most common problem targeted by data science in healthcare. About 30% of the selected papers deal with this dominant problem. Researchers may take into account the newest machine learning trends to enhance the performance of machine learning models and accelerate the development of models. Indeed, ensemble learning that combines the power of several individual machine-learning algorithms for a particular problem can provide better performance compared to the performance of an individual algorithm.

5. Related work

Different surveys and systematic reviews were proposed to assess and synthesize the published literature related to the application of data science and big data analytics to healthcare engineering systems, such as de la Torre Díez et al. (2016); Imran et al. (2020); Galetsi and Katsaliaki (2020a,b); Jiang et al. (2019); Khanra et al. (2020); Kruse et al. (2016); Pashazadeh and Navimipour (2018); Pramanik et al. (2020); Raja et al. (2020); Salazar-Reyna et al. (2020). The used systematic process, described in sub-section 2.2.1, allowed us to identify four secondary studies as most relevant ones, namely Galetsi and Katsaliaki (2020b); Khanra et al. (2020); Raja et al. (2020); Salazar-Reyna et al. (2020). As we are aware that other literature reviews may have appeared after we started this study, we applied a forward snowballing process on the already retained studies. Four new secondary studies published in 2022 and 2023 were identified, namely Etemadi et al. (2023); Khan et al. (2022); Miah et al. (2022) and Sabharwal and Miah (2022). Three among them, i.e. Etemadi et al. (2023); Khan et al. (2022); Sabharwal and Miah (2022) rejoin other eliminated studies such as Imran et al. (2020) because they only focus on a particular aspect of big data analytics in healthcare. As a matter of fact, Imran et al. (2020) provided a review of healthcare big data analytics applications dealing with NoSQL databases, Etemadi et al. (2023) studied healthcare recommender systems in systematic way, while Khan et al. (2022) and Sabharwal and Miah (2022) analyzed existing literature in the scope of a narrow healthcare application target, which is diagnosing diseases.

The relevant studies that may be compared with us are summarized in Table 2. All of them provided a comprehensive systematic review in order to determine to which extent big data analytics applications are adopted in healthcare. Salazar-Reyna et al. (2020) is the sole work that contextualize the research within data science. Raja et al. (2020) assessed 34 journal articles published between 2015 and 2019, and selected from 2 digital libraries (i.e. IEEE Xplore and Science Direct). Galetsi and Katsaliaki (2020b) conducted a meta-analysis review based on 804 papers published between 2000 and 2016, and identified from 2 digital libraries (i.e. Scopus and Web of Science). Salazar-Reyna et al. (2020) proposed a systematic literature review based on 576 publications (105 theoretical publications and 471 application publications) published between 2004 and 2019, and identified through 3 digital libraries (i.e. EBSCOhost, ProQuest and Scopus). Khanra et al. (2020) examined 41 studies published between 2013 and 2019, and identified using 4 digital libraries (i.e. PsycINFO, PubMed, Scopus and Web of Science). Miah et al. (2022) identified 2941 studies published between 2012 and 2019 from 1 digital library which is Scopus.

Although these works aimed at synthesizing prior works within the general scope of healthcare field, they haven't succeeded to give a comprehensive big picture of data science application in healthcare based on grounded evidence, built thanks to high quality publications. Indeed, all of them, except Khanra et al. (2020), haven't evaluated the quality of the documents identified through the selection process. On the other hand, all the studies either lack or do not set up a well-defined driving research question. Furthermore, they fail in covering the various facets (i.e. the What, Why and How) required to construct the big picture about applying data science in healthcare.

6. Threats to validity

In this paper, a comprehensive overview of data science within healthcare applications is presented. To achieve our goal, we elaborated a research protocol and conducted it following a well-defined process recommended by Petersen et al. (2015). We then answered the research questions and discussed the results. This work may be subject to a set of threats to validity, detailed in the sequel.

Table 2: Related works comparison

	Raja et al. (2020)	**Galetsi and Katsaliaki (2020b)**	**Salazar-Reyna et al. (2020)**	**Khanra et al. (2020)**	**Miah et al. (2022)**	**this study**
Digital library sources	2 (IEEE Xplore, Science Direct)	2 (Scopus, Web of Science)	3 (EBSCOhost, ProQuest and Scopus)	4 (PsycINFO, PubMed, Scopus, Web of Science)	1 (Scopus)	6 (ACM Digital Library, IEEE Xplore, Google Scholar, PubMed, Scopus and Science Direct)
Quality of studies	Journals without quality filtering	No quality reported	Journals without quality filtering	Yes	Yes without quality filtering	Yes
Period	2015 - 2019	2000 - 2016	2004 - 2019	2013 - 2019	2012 - 2019	2015 - 2019
Number of selected papers	34	804	576	41	2941	211
Question driving study	No	Yes (not well-defined)	Yes (not well-defined)	No	No	Yes
What view	Partially	Partially	Partially	No	No	Yes
Why view	Partially	Partially	Partially	Partially	Partially	Yes
How view	Partially	Yes	Yes	Partially	Partially	Yes

- *Research questions.* As mentioned above, the goal of this work is to give an overview of data science use in healthcare domain. Therefore, we defined a relatively general questions answering mainly the what, why and how. Nevertheless, it is always possible to identify more specific questions to go deeper with this research topic.

- *External validity.* This threat is related to external validity and is about the extent to which the results of this study could be generalized. It stems from the non representativity of the primary studies set to the research topic. We mitigated this threat by following the guidelines of a well-defined and well-known systematic process adapted from Petersen et al. (2015). The first identified set was based on the results of a systematic literature review performed with respect to a well-defined systematic process and published in high quality journal. The relevant papers set was then extended by forward snowballing in order to get a broader picture covering recent years (2019-2021) not taken into account in the source secondary study. Moreover, we based this study on only high quality scientific works published in high quality journals excluding conferences, workshops, and gray literature.

- *Internal validity.* This threat refers to the influence of extraneous variables on the design of the study. It is related to an incomplete relationship between findings, introducing systematic errors. We mitigated to this threat by following a well-defined process to define and answer the research questions (see Section 2, where, among others, the relationship between research questions and research goals is properly described).

- *Construct validity.* This threat is related to the validity of extracted data with respect to research questions. To address this type of threats, we performed the automated search on multiple digital libraries to avoid biases that may exist as a result of publisher policies as well as business concerns. Moreover, screening of resulted papers for the secondary study selection on the one hand, and screening of the identified papers for primary studies selection on the other hand were carried out by both authors in order to minimize potential subjectivity in the assessment criteria application. Finally, information extraction was performed by both researchers; all discrepancies were discussed to be resolved and to reach an agreement.

- *Conclusion validity.* This threat concerns the relationship between the extracted data and the obtained results; in other words, the reproducibility of this research study. Threats related to conclusion validity were mitigated by following the protocol of (Petersen et al., 2015). All process steps were performed based on this well-known and defined protocol. All conclusions were derived from extracted data by the help of tables and figures in order to avoid subjective interpretations of the results. Moreover, this work process has been formalized and documented, allowing other researchers to replicate this study.

7. Conclusion

The objective of this study was to assess and synthesize the published literature in the purpose of understanding and getting a comprehensive overview of data science in healthcare domain. To achieve this aim, we conducted a systematic review following a well-defined process. We identified 211 primary studies published between 2015 and 2021. This set was used to answer the diverse questions covering the three facets (i.e. What, Why and How) required to get the big picture of data science in healthcare.

The analysis of the primary studies showed that techniques based on artificial intelligence and especially machine learning are by far the most investigated ones to apply data science in healthcare domain. Furthermore, it allowed us to identify some gaps in the existing literature; we therefore proposed potential key future research directions on the utilization of data science in the healthcare domain. The main ones include (1) giving more interest to unstructured data as it increases at a rapid rate and it represents a valuable asset for healthcare researchers and practitioners; and (2) investigating techniques allowing artificial wisdom in order to enable systems to understand the whole environment and become adoptable for a wise healthcare oriented decision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdullah, S.S., Rostamzadeh, N., Sedig, K., Garg, A.X., McArthur, E., 2020. Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records. Informatics 7, 1–30. doi:10.3390/informatics7020017.

Afzal, M., Hussain, M., Khan, W.A., Ali, T., Lee, S., Huh, E.N., Ahmad, H.F., Jamshed, A., Iqbal, H., Irfan, M., et al., 2017. Comprehensible knowledge model creation for cancer treatment decision making. Computers in Biology and Medicine 82, 119–129. doi:10.1016/j.compbiomed.2017.01.010.

Alakus, T.B., Turkoglu, I., 2020. Comparison of deep learning approaches to predict covid-19 infection. Chaos, Solitons & Fractals 140, 110120. doi:10.1016/j.chaos.2020.110120.

Alfian, G., Syafrudin, M., Anshari, M., Benes, F., Atmaji, F.T.D., Fahrurrozi, I., Hidayatullah, A.F., Rhee, J., 2020. Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features. Biocybernetics and Biomedical Engineering 40, 1586–1599. doi:10.1016/j.bbe.2020.10.004.

Alhamid, M.F., 2019. Investigation of mammograms in the cloud for smart healthcare. Multimedia Tools and Applications 78, 8997–9009. doi:https://doi.org/10.1007/s11042-017-5239-z.

Alharithi, F., Almulihi, A., Bourouis, S., Alroobaea, R., Bouguila, N., 2021. Discriminative learning approach based on flexible mixture model for medical data categorization and recognition. Sensors 21, 2450. doi:10.3390/s21072450.

Alharthi, R., Alharthi, R., Guthier, B., ElSaddik, A., 2019. Casp: context-aware stress prediction system. Multimedia Tools and Applications 78, 9011–9031. doi:10.1007/s11042-017-5246-0.

Alhussein, M., Muhammad, G., 2019. Automatic voice pathology monitoring using parallel deep models for smart healthcare. Ieee Access 7, 46474–46479. doi:10.1109/access.2019.2905597.

Almasoud, A.M., Al-Khalifa, H.S., Al-Salman, A.S., 2019. Handling big data scalability in biological domain using parallel and distributed processing: a case of three biological semantic similarity measures. BioMed Research International 2019. doi:10.1155/2019/6750296.

Alshakhs, F., Alharthi, H., Aslam, N., Khan, I.U., Elasheri, M., 2020. Predicting postoperative length of stay for isolated coronary artery bypass graft patients using machine learning. International Journal of General Medicine , 751–762doi:10.2147/IJGM.S250334.

Arefeen, M.A., Nimi, S.T., Rahman, M.S., Arshad, S.H., Holloway, J.W., Rezwan, F.I., 2020. Prediction of lung function in adolescence using epigenetic aging: a machine learning approach. Methods and Protocols 3, 77. doi:10.3390/mps3040077.

Ashfaq, A., Sant'Anna, A., Lingman, M., Nowaczyk, S., 2019. Readmission prediction using deep learning on electronic health records. Journal of biomedical informatics 97, 103256. doi:10.1016/j.jbi.2019.103256.

Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H., 2018. Improving palliative care with deep learning. BMC medical informatics and decision making 18, 55–64. doi:10.1186/s12911-018-0677-8.

Baechle, C., Agarwal, A., 2017. A framework for the estimation and reduction of hospital readmission penalties using predictive analytics. Journal of Big Data 4, 1–15. doi:10.1186/s40537-017-0098-z.

Batool, H., Usman Akram, M., Batool, F., Butt, W.H., 2016. Intelligent framework for diagnosis of frozen shoulder using cross sectional survey and case studies. SpringerPlus 5, 1048. doi:10.1186/s40064-016-3537-y.

Baytas, I.M., Lin, K., Wang, F., Jain, A.K., Zhou, J., 2016. Phenotree: Interactive visual analytics for hierarchical phenotyping from large-scale electronic health records. IEEE Transactions on Multimedia 18, 2257–2270. doi:10.1109/TMM.2016.2614225.

Beeksma, M., Verberne, S., van den Bosch, A., Das, E., Hendrickx, I., Groenewoud, S., 2019. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC medical informatics and decision making 19, 1–15. doi:10.1186/s12911-019-0775-2.

Bergquist, T., Yan, Y., Schaffter, T., Yu, T., Pejaver, V., Hammarlund, N., Prosser, J., Guinney, J., Mooney, S., 2020. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. Journal of the American Medical Informatics Association 27, 1393–1400. doi:10.1093/jamia/ocaa083.

Blom, M.C., Ashfaq, A., Sant'Anna, A., Anderson, P.D., Lingman, M., 2019. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study. BMJ open 9, e028015. doi:10.1136/bmjopen-2018-028015.

Cao, K., Verspoor, K., Sahebjada, S., Baird, P.N., 2020. Evaluating the performance of various machine learning algorithms to detect subclinical keratoconus. Translational Vision Science & Technology 9, 24–24. doi:10.1167/tvst.9.2.24.

Cao, L., 2016. Data science: nature and pitfalls. IEEE Intelligent Systems 31, 66–75. doi:10.1109/mis.2016.86.

Cao, L., 2017a. Data science: a comprehensive overview. ACM Computing Surveys (CSUR) 50, 1–42. doi:https://doi.org/10.1145/3076253.

Cao, L., 2017b. Data science: challenges and directions. Communications of the ACM 60, 59–68. doi:10.1145/3015456.

Castiglione, A., Vijayakumar, P., Nappi, M., Sadiq, S., Umer, M., 2021. Covid-19: automatic detection of the novel coronavirus disease from ct images using an optimized convolutional neural network. IEEE Transactions on Industrial Informatics 17, 6480–6488. doi:10.1109/tii.2021.3057524.

Chang, V., 2018. Computational intelligence for medical imaging simulations. Journal of medical systems 42, 1–12. doi:https://doi.org/10.1007/s10916-017-0861-x.

Chang, W., Ji, X., Xiao, Y., Zhang, Y., Chen, B., Liu, H., Zhou, S., 2021. Prediction of hypertension outcomes based on gain sequence forward tabu search feature selection and xgboost. Diagnostics 11, 792. doi:10.3390/diagnostics11050792.

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., Zhou, S., 2019. A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics 9, 178. doi:10.3390/diagnostics9040178.

Chao-Wen, C., Yuh-Wen, C., Moussa, L., Tzung-Hung, L., 2015. Using multi-objective affinity model for mining the rules of revisits within 72 hours for emergency department patients. Multiple Criteria Decision Making , 5–31.

Chen, S., Wu, S., 2020. Deep learning for identifying environmental risk factors of acute respiratory diseases in beijing, china: implications for population with different age and gender. International Journal of Environmental Health Research 30, 435–446. doi:10.1080/09603123.2019.1597836.

Chen, V.C.H., Lin, T.Y., Yeh, D.C., Chai, J.W., Weng, J.C., 2019. Predicting chemo-brain in breast cancer survivors using multiple mri features and machine-learning. Magnetic Resonance in Medicine 81, 3304–3313. doi:https://doi.org/10.1002/mrm.27607.

Cho, I.J., Sung, J.M., Kim, H.C., Lee, S.E., Chae, M.H., Kavousi, M., Rueda-Ochoa, O.L., Ikram, M.A., Franco, O.H., Min, J.K., et al., 2019. Development and external validation of a deep learning algorithm for prognostication of cardiovascular outcomes. Korean circulation journal 50, 72–84. doi:10.4070/kcj.2019.0105.

Cho, S.E., Geem, Z.W., Na, K.S., 2020. Prediction of suicide among 372,813 individuals under medical check-up. Journal of psychiatric research 131, 9–14. doi:10.1016/j.jpsychires.2020.08.035.

Cho, S.E., Geem, Z.W., Na, K.S., 2021. Development of a suicide prediction model for the elderly using health screening data. International journal of environmental research and public health 18, 10150. doi:10.3390/ijerph181910150.

Choi, J., Choi, C., Ko, H., Kim, P., 2016. Intelligent healthcare service using health lifelog analysis. Journal of Medical Systems 40, 1–10. doi:10.1007/s10916-016-0534-1.

Choi, Y., Boo, Y., 2020. Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. International journal of environmental research and public health 17, 897. doi:10.3390/ijerph17030897.

Cielen, D., Meysman, A., 2016. Introducing data science: big data, machine learning, and more, using Python tools. Simon and Schuster.

Connolly, B., Cohen, K.B., Santel, D., Bayram, U., Pestian, J., 2017. A nonparametric bayesian method of translating machine learning scores to probabilities in clinical decision support. BMC bioinformatics 18, 1–12. doi:10.1186/s12859-017-1736-3.

Crump, C.A., Wernz, C., Schlachta-Fairchild, L., Steidle, E., Duncan, A., Cathers, L., 2021. Closing the digital health evidence gap: development of a predictive score to maximize patient outcomes. Telemedicine and e-Health 27, 1029–1038. doi:10.1089/tmj.2020.0334.

Datta, A., Matlock, M.K., Le Dang, N., Moulin, T., Woeltje, K.F., Yanik, E.L., Swamidass, S.J., 2020. ‘black box’to ‘conversational’machine learning: Ondansetron reduces risk of hospital-acquired venous thromboembolism. IEEE Journal of Biomedical and Health Informatics 25, 2204–2214. doi:10.1109/jbhi.2020.3033405.

Djatna, T., Hardhienata, M.K.D., Masruriyah, A.F.N., 2018. An intuitionistic fuzzy diagnosis analytics for stroke disease. Journal of Big Data 5, 1–14. doi:10.1186/s40537-018-0142-7.

Dorr, B.J., Greenberg, C.S., Fontana, P.C., Przybocki, M.A., Bras, M.L., Ploehn, C.A., Aulov, O., Michel, M., Golden, E.J., Chang, W., 2016. A new data science research program: evaluation, metrology, standards, and community outreach. International Journal of Data Science and Analytics 1, 177–197. doi:10.1007/s41060-016-0016-z.

El Khatib, M., Hamidi, S., Al Ameeri, I., Al Zaabi, H., Al Marqab, R., 2022. Digital disruption and big data in healthcare-opportunities and challenges. ClinicoEconomics and Outcomes Research , 563–574.

Elakkiya, R., Vijayakumar, P., Karuppiah, M., 2021. Covid_screenet: Covid-19 screening in chest radiography images using deep transfer stacking. Information Systems Frontiers 23, 1369–1383. doi:10.1007/s10796-021-10123-x.

Etemadi, M., Bazzaz Abkenar, S., Ahmadzadeh, A., Haghi Kashani, M., Asghari, P., Akbari, M., Mahdipour, E., 2023. A systematic review of healthcare recommender systems: Open issues, challenges, and techniques. Expert Systems with Applications 213, 118823. doi:https://doi.org/10.1016/j.eswa.2022.118823.

Eyuboglu, S., Angus, G., Patel, B.N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., Lungren, M.P., 2021. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body fdg-pet/ct. Nature communications 12, 1–15. doi:10.1038/s41467-021-22018-1.

Fairie, P., Zhang, Z., D'Souza, A.G., Walsh, T., Quan, H., Santana, M.J., 2021. Categorising patient concerns using natural language processing techniques. BMJ health & care informatics 28. doi:10.1136/bmjhci-2020-100274.

Feldman, K., Davis, D., Chawla, N.V., 2015. Scaling and contextualizing personalized healthcare: A case study of disease prediction algorithm integration. Journal of biomedical informatics 57, 377–385. doi:10.1016/j.jbi.2015.07.017.

Fernández Del Carpio, A., Angarita, L.B., 2018. Techniques based on data science for software processes: A systematic literature review, in: Software Process Improvement and Capability Determination, Springer International Publishing, Cham. pp. 16–30. doi:10.1007/978-3-030-00623-5_2.

Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., Song, M., 2019. Mining of textual health information from reddit: Analysis of chronic diseases with extracted entities and their relations. Journal of medical Internet research 21, e12876. doi:10.2196/12876.

Frické, M.H., 2018. Data-information-knowledge-wisdom (dikw) pyramid, framework, continuum. Encyclopedia of Big Data , 1–4doi:10.1007/978-3-319-32010-6_331.

Fritzell, P., Mesterton, J., Hagg, O., 2021. Prediction of outcome after spinal surgery—using the dialogue support based on the swedish national quality register. European Spine Journal , 1–12doi:10.1007/s00586-021-07065-y.

Galetsi, P., Katsaliaki, K., 2020a. A review of the literature on big data analytics in healthcare. Journal of the Operational Research Society 71, 1511–1529. doi:10.1080/01605682.2019.1630328.

Galetsi, P., Katsaliaki, K., 2020b. A review of the literature on big data analytics in healthcare. Journal of the Operational Research Society 71, 1511–1529. doi:10.1080/01605682.2019.1630328.

Gálvez, J.A., Jalali, A., Ahumada, L., Simpao, A.F., Rehman, M.A., 2017. Neural network classifier for automatic detection of invasive versus noninvasive airway management technique based on respiratory monitoring parameters in a pediatric anesthesia. Journal of medical systems 41, 153. doi:10.1007/s10916-017-0787-3.

Ganggayah, M.D., Taib, N.A., Har, Y.C., Lio, P., Dhillon, S.K., 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making 19, 1–17. doi:10.1186/s12911-019-0801-4.

Golas, S.B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., Hisamitsu, T., Kojima, G., Felsted, J., Kakarmath, S., et al., 2018. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC medical informatics and decision making 18, 1–17. doi:10.1186/s12911-018-0620-z.

Goto, T., Jo, T., Matsui, H., Fushimi, K., Hayashi, H., Yasunaga, H., 2019. Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. COPD: Journal of Chronic Obstructive Pulmonary Disease 16, 338–343. doi:10.1080/15412555.2019.1688278.

Group, B., . Big data taxonomy. Technical Report. URL: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf.

Guo, A., Foraker, R., White, P., Chivers, C., Courtright, K., Moore, N., 2021. Using electronic health records and claims data to identify high-risk patients likely to benefit from palliative care. American Journal of Managed Care 27.

Guo, L., Jin, B., Yao, C., Yang, H., Huang, D., Wang, F., et al., 2016. Which doctor to trust: a recommender system for identifying the right doctors. Journal of medical Internet research 18, e6015. doi:10.2196/jmir.6015.

Han, S.H., Kim, K.O., Cha, E.J., Kim, K.A., Shon, H.S., 2017. System framework for cardiovascular disease prediction based on big data technology. Symmetry 9, 293. doi:10.3390/sym9120293.

Hao, H., Zhang, K., 2016. The voice of chinese health consumers: a text mining approach to web-based physician reviews. Journal of medical Internet research 18, e108. doi:10.2196/jmir.4430.

Harar, P., Galaz, Z., Alonso-Hernandez, J.B., Mekyska, J., Burget, R., Smekal, Z., 2020. Towards robust voice pathology detection. Neural Computing and Applications 32, 15747–15757. doi:https://doi.org/10.1007/s00521-018-3464-7.

Hossain, M.S., Muhammad, G., 2016. Healthcare big data voice pathology assessment framework. IEEE Access 4, 7806–7815. doi:10.1109/access.2016.2626316.

Hu, J., Zhang, X., Yang, Y., Liu, Y., Chen, X., 2020. New doctors ranking system based on vikor method. International Transactions in Operational Research 27, 1236–1261. doi:https://doi.org/10.1111/itor.12569.

Hu, S., Teng, F., Huang, L., Yan, J., Zhang, H., 2021. An explainable cnn approach for medical codes prediction from clinical text. BMC Medical Informatics and Decision Making 21, 1–12. doi:10.1186/s12911-021-01615-6.

Huang, M., Chen, Y., Chen, B.W., Liu, J., Rho, S., Ji, W., 2016. A semi-supervised privacy-preserving clustering algorithm for healthcare. Peer-to-Peer Networking and Applications 9, 864–875. doi:10.1007/s12083-015-0356-9.

Hyland, S.L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, A., Ratsch, G., Merz, T.M., 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine 26, 364–373. doi:https://doi.org/10.1038/s41591-020-0789-4.

Imran, S., Mahmood, T., Morshed, A., Sellis, T., 2020. Big data analytics in healthcare - a systematic literature review and roadmap for practical implementation. IEEE/CAA Journal of Automatica Sinica 8, 1–22. doi:10.1109/jas.2020.1003384.

Jain, P., Agarwal, A., Behara, R., Baechle, C., 2019. HPCC based framework for COPD readmission risk analysis. Journal of Big Data 6, 1–13. doi:10.1186/s40537-019-0189-0.

Jayalakshmi, M., Rao, S.N., 2020. Discrete wavelet transmission and modified pso with aco based feed forward neural network model for brain tumour detection. Computers, Materials & Continua 65, 1081–1096. doi:10.32604/cmc.2020.011710.

Jiang, Y., Luo, Z., Wang, Z., Lin, B., 2019. Review of thermal comfort infused with the latest big data and modeling progresses in public health. Building and Environment 164, 106336. doi:10.1016/j.buildenv.2019.106336.

Jifa, G., 2013. Data, information, knowledge, wisdom and meta-synthesis of wisdom-comment on wisdom global and wisdom cities. Procedia Computer Science 17, 713–719. doi:10.1016/j.procs.2013.05.092.

Kadiri, S.R., Alku, P., 2019. Analysis and detection of pathological voice using glottal source features. IEEE Journal of Selected Topics in Signal Processing 14, 367–379. doi:10.1109/jstsp.2019.2957988.

Kale, V.V., Hamde, S.T., Holambe, R.S., 2019. Multi class disorder detection of magnetic resonance brain images using composite features and neural network. Biomedical engineering letters 9, 221–231. doi:10.1007/s13534-019-00103-1.

Kanegae, H., Suzuki, K., Fukatani, K., Ito, T., Harada, N., Kario, K., 2020. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. The Journal of Clinical Hypertension 22, 445–450. doi:https://doi.org/10.1111/jch.13759.

Khan, S., Khan, H.U., Nazir, S., 2022. Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing. Scientific Reports 12, 22377. doi:10.1038/s41598-022-26090-5.

Khanra, S., Dhir, A., Islam, A.N., Mäntymäki, M., 2020. Big data analytics in healthcare: a systematic literature review. Enterprise Information Systems 14, 878–912. doi:10.1080/17517575.2020.1812005.

Khasha, R., Sepehri, M.M., Mahdaviani, S.A., 2019. An ensemble learning method for asthma control level detection with leveraging medical knowledge-based classifier and supervised learning. Journal of medical systems 43, 1–15. doi:10.1007/s10916-019-1259-8.

Khasha, R., Sepehri, M.M., Taherkhani, N., 2021. Detecting asthma control level using feature-based time series classification. Applied Soft Computing 111, 107694. doi:10.1016/j.asoc.2021.107694.

Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., Linkman, S., 2010. Systematic literature reviews in software engineering – a tertiary study. Information and Software Technology 52, 792–805. doi:10.1016/j.infsof.2010.03.006.

Krempel, R., Kulkarni, P., Yim, A., Lang, U., Habermann, B., Frommolt, P., 2018. Integrative analysis and machine learning on cancer genomics data using the cancer systems biology database (cancersysdb). BMC bioinformatics 19, 1–10. doi:10.1186/s12859-018-2157-7.

Kruse, C.S., Goswamy, R., Raval, Y.J., Marawi, S., 2016. Challenges and opportunities of big data in health care: a systematic review. JMIR medical informatics 4, e5359. doi:10.2196/medinform.5359.

Li, L., Zhao, J., Hou, L., Zhai, Y., Shi, J., Cui, F., 2019a. An attention-based deep learning model for clinical named entity recognition of chinese electronic medical records. BMC Medical Informatics and Decision Making 19, 1–11. doi:10.1186/s12911-019-0933-6.

Li, W., Song, Y., Chen, K., Ying, J., Zheng, Z., Qiao, S., Yang, M., Zhang, M., Zhang, Y., 2021. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in china. BMJ open 11, e050989. doi:10.1136/bmjopen-2021-050989.

Li, X., Wang, H., He, H., Du, J., Chen, J., Wu, J., 2019b. Intelligent diagnosis with chinese electronic medical records based on convolutional neural networks. BMC bioinformatics 20, 1–12. doi:10.1186/s12859-019-2617-8.

Liu, H., Wang, X., Tang, K., Peng, E., Xia, D., Chen, Z., 2021. Machine learning-assisted decision-support models to better predict patients with calculous pyonephrosis. Translational Andrology and Urology 10, 710. doi:10.21037/tau-20-1208.

Liu, J., Bier, E., Wilson, A., Guerra-Gomez, J.A., Honda, T., Sricharan, K., Gilpin, L., Davies, D., 2016. Graph analysis for detecting fraud, waste, and abuse in healthcare data. Ai Magazine 37, 33–46. doi:10.1609/aimag.v37i2.2630.

Liu, X., Lu, R., Ma, J., Chen, L., Qin, B., 2015. Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. IEEE journal of biomedical and health informatics 20, 655–668. doi:10.1109/jbhi.2015.2407157.

Lopez Bernal, S., Martinez Valverde, J., Huertas Celdran, A., Martinez Perez, G., 2021a. SENIOR: An intelligent web-based ecosystem to predict high blood pressure adverse events using biomarkers and environmental data. Applied Sciences 11, 2506. doi:10.3390/app11062506.

Lopez Bernal, S., Martinez Valverde, J., Huertas Celdran, A., Martinez Perez, G., 2021b. Senior: An intelligent web-based ecosystem to predict high blood pressure adverse events using biomarkers and environmental data. Applied Sciences 11, 2506. doi:10.3390/app11062506.

Lu, H., Uddin, S., 2021. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. Scientific reports 11, 22607. doi:10.1038/s41598-021-01964-2.

Lu, X.H., Liu, A., Fuh, S.C., Lian, Y., Guo, L., Yang, Y., Marelli, A., Li, Y., 2021. Recurrent disease progression networks for modelling risk trajectory of heart failure. PloS one 16, e0245177. doi:10.1371/journal.pone.0245177.

MacKay, E.J., Stubna, M.D., Chivers, C., Draugelis, M.E., Hanson, W.J., Desai, N.D., Groeneveld, P.W., 2021. Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. PloS one 16, e0252585. doi:10.1371/journal.pone.0252585.

Makino, M., Yoshimoto, R., Ono, M., Itoko, T., Katsuki, T., Koseki, A., Kudo, M., Haida, K., Kuroda, J., Yanagiya, R., et al., 2019. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Scientific reports 9, 11862. doi:10.1038/s41598-019-48263-5.

Massaro, A., Maritati, V., Giannone, D., Convertini, D., Galiano, A., 2019. Lstm dss automatism and dataset optimization for diabetes prediction. Applied Sciences 9, 3532. doi:10.3390/app9173532.

Miah, S.J., Camilleri, E., Vu, H.Q., 2022. Big data in healthcare research: a survey study. Journal of Computer Information Systems 62, 480–492. doi:10.1080/08874417.2020.1858727.

Murphy, D.R., Meyer, A.N.D., Sittig, D.F., Meeks, D.W., Thomas, E.J., Singh, H., 2019. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Quality & Safety 28, 151–159.

Nambiar, R., Bhardwaj, R., Sethi, A., Vargheese, R., 2013. A look at challenges and opportunities of big data analytics in healthcare, in: 2013 IEEE international conference on Big Data, IEEE. pp. 17–22. doi:10.1109/bigdata.2013.6691753.

Narayan, S., Sathiyamoorthy, E., 2019. A novel recommender system based on fft with machine learning for predicting and identifying heart diseases. Neural Computing and Applications 31, 93–102. doi:10.1007/s00521-018-3662-3.

Nasir, M., South-Winter, C., Ragothaman, S., Dag, A., 2018. A comparative data analytic approach to construct a risk trade-off for cardiac patients' re-admissions. Industrial Management & Data Systems doi:10.1108/imds-12-2017-0579.

Nayan, M., Salari, K., Bozzo, A., Ganglberger, W., Carvalho, F., Feldman, A.S., Trinh, Q.D., 2021. Predicting survival after radical prostatectomy: Variation of machine learning performance by race. The Prostate 81, 1355–1364. doi:10.1002/pros.24233.

Nayan, M., Salari, K., Bozzo, A., Ganglberger, W., Lu, G., Carvalho, F., Gusev, A., Schneider, A., Westover, B.M., Feldman, A.S., 2022. A machine learning approach to predict progression on active surveillance for prostate cancer. Urologic Oncology: Seminars and Original Investigations 40, 161.e1–161.e7. doi:https://doi.org/10.1016/j.urolonc.2021.08.007.

Nicholson, N.C., Giusti, F., Bettio, M., Negrao Carvalho, R., Dimitrova, N., Dyba, T., Flego, M., Neamtiu, L., Randi, G., Martos, C., 2021. An ontology to model the international rules for multiple primary malignant tumours in cancer registration. Applied Sciences 11, 7233. doi:10.3390/app11167233.

Nielsen, O.B., 2017. A comprehensive review of data governance literature. Selected Papers IRIS 8, 120–133.

Ning, X., Fan, Z., Burgun, E., Ren, Z., Schleyer, T., 2021. Improving information retrieval from electronic health records using dynamic and multi-collaborative filtering. Plos one 16, e0255467. doi:10.1371/journal.pone.0255467.

Nozais, V., Boutinaud, P., Verrecchia, V., Gueye, M.F., Hervé, P.Y., Tzourio, C., Mazoyer, B., Joliot, M., 2021. Deep learning-based classification of resting-state fmri independent-component analysis. Neuroinformatics 19, 619–637. doi:10.1007/s12021-021-09514-x.

Nudel, J., Bishara, A.M., de Geus, S.W.L., Patil, P., Srinivasan, J., Hess, D.T., Woodson, J., 2021. Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the mbsaqip database. Surgical endoscopy 35, 182–191. doi:10.1007/s00464-020-07378-x.

OECD, 2021. OECD Health Statistics. Technical Report. URL: https://www.oecd.org/els/healthsystems/health-data.htm.

Ontotext, . What is the data, information, knowledge, wisdom (dikw) pyramid? URL: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/.

Pashazadeh, A., Navimipour, N.J., 2018. Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. Journal of biomedical informatics 82, 47–62. doi:10.1016/j.jbi.2018.03.014.

Patil, D., 2012. Data Jujitsu: The Art of Turning Data into Product. O'Reilly Media, Inc.

Pendleton, S.C., Slater, L.T., Karwath, A., Gilbert, R.M., Davis, N., Pesudovs, K., Liu, X., Denniston, A.K., Gkoutos, G.V., Braithwaite, T., 2021. Development and application of the ocular immune-mediated inflammatory diseases ontology enhanced with synonyms from online patient support forum conversation. Computers in biology and medicine 135, 104542. doi:10.1016/j.compbiomed.2021.104542.

Periasamy, K., Periasamy, S., Velayutham, S., Zhang, Z., Ahmed, S.T., Jayapalan, A., 2022. A proactive model to predict osteoporosis: An artificial immune system approach. Expert Systems 39, e12708. doi:10.1111/exsy.12708.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M., 2008. Systematic mapping studies in software engineering, in: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, pp. 1–10. doi:10.14236/ewic/ease2008.8.

Petersen, K., Vakkalanka, S., Kuzniarz, L., 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology 64, 1–18. doi:10.1016/j.infsof.2015.03.007.

Polce, E.M., Kunze, K.N., Fu, M.C., Garrigues, G.E., Forsythe, B., Nicholson, G.P., Cole, B.J., Verma, N.N., 2021. Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. Journal of Shoulder and Elbow Surgery 30, e290–e299. doi:10.1016/j.jse.2020.09.007.

Pramanik, M.I., Lau, R.Y., Azad, M.A.K., Hossain, M.S., Chowdhury, M.K.H., Karmaker, B., 2020. Healthcare informatics and analytics in big data. Expert Systems with Applications 152, 113388. doi:https://doi.org/10.1016/j.eswa.2020.113388.

Probyto, 2020. Data Science for Business Professionals: A practicle guide for beginners. Manish Jain for BPB Publications.

Pustišek, M., 2017. A system for multi-domain contextualization of personal health data. Journal of medical systems 41, 1–6. doi:https://doi.org/10.1007/s10916-016-0663-6.

Rabkin, M., 2016. Data, information, intelligence, knowledge and wisdom. URL: https://www.linkedin.com/pulse/data-information-intelligence-knowledge-wisdom-mark-rabkin.

Raghavaiah, P., Varadarajan, S., 2021. A cad system design to diagnosize alzheimers disease from mri brain images using optimal deep neural network. Multimedia Tools and Applications 80, 26411–26428. doi:10.1007/s11042-021-10928-7.

Raja, R., Mukherjee, I., Sarkar, B.K., 2020. A systematic review of healthcare big data. Scientific programming 2020. doi:10.1155/2020/5471849.

Razavi, F., Tarokh, M.J., Alborzi, M., 2019. An intelligent alzheimer's disease diagnosis method using unsupervised feature learning. Journal of Big Data 6, 1–16. doi:10.1186/s40537-019-0190-7.

Richter, A.N., Khoshgoftaar, T.M., 2019a. Efficient learning from big data for cancer risk modeling: a case study with melanoma. Computers in biology and medicine 110, 29–39. doi:10.1016/j.compbiomed.2019.04.039.

Richter, A.N., Khoshgoftaar, T.M., 2019b. Efficient learning from big data for cancer risk modeling: a case study with melanoma. Computers in biology and medicine 110, 29–39. doi:10.1016/j.compbiomed.2019.04.039.

Ross, E.G., Jung, K., Dudley, J.T., Li, L., Leeper, N.J., Shah, N.H., 2019. Predicting future cardiovascular events in patients with peripheral artery disease using electronic health record data. Circulation: Cardiovascular Quality and Outcomes 12, e004741. doi:10.1161/circoutcomes.118.004741.

Ross, E.G., Shah, N.H., Dalman, R.L., Nead, K.T., Cooke, J.P., Leeper, N.J., 2016. The use of machine learning for the identification of peripheral artery disease and future mortality risk. Journal of vascular surgery 64, 1515–1522. doi:10.1016/j.jvs.2016.04.026.

Rowley, J., 2007. The wisdom hierarchy: representations of the dikw hierarchy. Journal of information science 33, 163–180. doi:10.1177/0165551506070706.

Sabharwal, R., Miah, S.J., 2022. An intelligent literature review: adopting inductive approach to define machine learning applications in the clinical domain. Journal of Big Data 9, 1–18. doi:10.1186/s40537-022-00605-3.

Sadoughi, F., Behmanesh, A., Sayfouri, N., 2020. Internet of things in medicine: A systematic mapping study. Journal of Biomedical Informatics 103, 103383. doi:10.1016/j.jbi.2020.103383.

Sadrawi, M., Lin, Y.T., Lin, C.H., Mathunjwa, B., Hsin, H.T., Fan, S.Z., Abbod, M.F., Shieh, J.S., 2021. Non-invasive hemodynamics monitoring system based on electrocardiography via deep convolutional autoencoder. Sensors 21, 6264. doi:10.3390/s21186264.

Sadrawi, M., Sun, W.Z., Ma, M.H.M., Yeh, Y.T., Abbod, M.F., Shieh, J.S., 2018. Ensemble genetic fuzzy neuro model applied for the emergency medical service via unbalanced data evaluation. Symmetry 10, 71. doi:10.3390/sym10030071.

Sahni, N., Tourani, R., Sullivan, D., Simon, G., 2020. min-sia: a lightweight algorithm to predict the risk of 6-month mortality at the time of hospital admission. Journal of general internal medicine 35, 1413–1418. doi:10.1007/s11606-020-05733-1.

Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E.M., Diaz-Ramirez, J., Garza-Reyes, J.A., Kumar, A., 2020. A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. Management Decision doi:10.1108/md-01-2020-0035.

Saranya, P., Prabakaran, S., 2020. Automatic detection of non-proliferative diabetic retinopathy in retinal fundus images using convolution neural network. Journal of Ambient Intelligence and Humanized Computing , 1–10doi:10.1007/s12652-020-02518-6.

Sena, G.R., Lima, T.P.F., Mello, M.J.G., Thuler, L.C.S., Lima, J.T.O., 2019. Developing machine learning algorithms for the prediction of early death in elderly cancer patients: usability study. JMIR cancer 5, e12163. doi:10.2196/12163.

Shah, A.M., Yan, X., Shah, S.A.A., Mamirkulova, G., 2020. Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. Journal of Ambient Intelligence and Humanized Computing 11, 2925–2942. doi:10.1007/s1265.

Shanker, R., Bhattacharya, M., 2021. Automated diagnosis system for detection of the pathological brain using fast version of simplified pulse-coupled neural network and twin support vector machine. Multimedia Tools and Applications 80, 30479–30502. doi:10.1007/s11042-021-10937-6.

Shen, Z., Wu, Q., Wang, Z., Chen, G., Lin, B., 2021. Diabetic retinopathy prediction by ensemble learning based on biochemical and physical data. Sensors 21, 3663. doi:10.3390/s21113663.

Singh, P., Feder, S., Jones, M., Meyer, A., 2021. Market Guide for Digital Health Platform for Healthcare Providers. Technical Report. Gartner. URL: https://www.gartner.com/en/documents/4006616.

Srinivasu, P.N., SivaSai, J.G., Ijaz, M.F., Bhoi, A.K., Kim, W., Kang, J.J., 2021. Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm. Sensors 21, 2852.

Tahmassebi, A., Gandomi, A.H., Schulte, M.H., Goudriaan, A.E., Foo, S.Y., Meyer-Baese, A., 2018. Optimized naive-bayes and decision tree approaches for fmri smoking cessation classification. Complexity 2018. doi:10.1155/2018/2740817.

Tai, Y., Qian, K., Huang, X., Zhang, J., Jan, M.A., Yu, Z., 2021. Intelligent intraoperative haptic-ar navigation for covid-19 lung biopsy using deep hybrid model. IEEE Transactions on Industrial Informatics 17, 6519–6527. doi:10.1109/tii.2021.3052788.

Tomašev, N., Harris, N., Baur, S., Mottram, A., Glorot, X., Rae, J.W., Zielinski, M., Askham, H., Saraiva, A., Magliulo, V., et al., 2021. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nature Protocols 16, 2765–2787. doi:10.1038/s41596-021-00513-5.

de la Torre Díez, I., Cosgaya, H.M., Garcia-Zapirain, B., López-Coronado, M., 2016. Big data in health: a literature review from the year 2005. Journal of medical systems 40, 1–6. doi:10.1007/s10916-016-0565-7.

Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C., Price, D., 2016. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. Artificial intelligence in medicine 74, 44–52. doi:10.1016/j.artmed.2016.11.003.

Tsai, C.W., Chiang, M.C., Ksentini, A., Chen, M., 2016. Metaheuristic algorithms for healthcare: Open issues and challenges. Computers & Electrical Engineering 53, 421–434. doi:10.1016/j.compeleceng.2016.03.005.

Ulloa Cerna, A.E., Jing, L., Good, C.W., vanMaanen, D.P., Raghunath, S., Suever, J.D., Nevius, C.D., Wehner, G.J., Hartzel, D.N., Leader, J.B., et al., 2021. Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality. Nature Biomedical Engineering 5, 546–554. doi:10.1038/s41551-020-00667-9.

Verde, L., De Pietro, G., Alrashoud, M., Ghoneim, A., Al-Mutib, K.N., Sannino, G., 2019. Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app. IEEE Access 7, 124048–124054. doi:10.1109/access.2019.2938265.

Veroneze, R., Cruz Tfaile Corbi, S., Roque da Silva, B., de S. Rocha, C., V. Maurer-Morelli, C., Perez Orrico, S.R., Cirelli, J.A., Von Zuben, F.J., Mantuaneli Scarel-Caminaga, R., 2020. Using association rule mining to jointly detect clinical features and differentially expressed genes related to chronic inflammatory diseases. PloS one 15, e0240269. doi:10.1371/journal.pone.0240269.

Wang, T., Lu, C., Shen, G., 2019. Detection of sleep apnea from single-lead ecg signal using a time window artificial neural network. BioMed research international 2019. doi:10.1155/2019/9768072.

Wang, Z., Chen, L., Zhang, J., Yin, Y., Li, D., 2020. Multi-view ensemble learning with empirical kernel for heart failure mortality prediction. International journal for numerical methods in biomedical engineering 36, e3273. doi:https://doi.org/10.1002/cnm.3273.

Waqar, M., Majeed, N., Dawood, H., Daud, A., Aljohani, N.R., 2019. An adaptive doctor-recommender system. Behaviour & Information Technology 38, 959–973. doi:10.1080/0144929X.2019.1625441.

Weegar, R., Sundström, K., 2020. Using machine learning for predicting cervical cancer from swedish electronic health records by mining hierarchical representations. PloS one 15, e0237911. doi:10.1371/journal.pone.0237911.

Wimmer, H., Yoon, V.Y., Sugumaran, V., 2016. A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy. Decision Support Systems 88, 51–66. doi:10.1016/j.dss.2016.05.008.

Wong, B., Ho, G.T., Tsui, E., 2017. Development of an intelligent e-healthcare system for the domestic care industry. Industrial Management & Data Systems 117, 1426–1445. doi:10.1108/imds-08-2016-0342.

Wong, N.C., Lam, C., Patterson, L., Shayegan, B., 2019. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. BJU international 123, 51–57. doi:10.1111/bju.14477.

Woo, H., Kim, K., Cha, K., Lee, J.Y., Mun, H., Cho, S.J., Chung, J.I., Pyo, J.H., Lee, K.C., Kang, M., et al., 2019. Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. Journal of medical Internet research 21, e10013. doi:10.2196/10013.

Wu, J., Qin, S., Wang, J., Li, J., Wang, H., Li, H., Chen, Z., Li, C., Wang, J., Yuan, J., 2020. Develop and evaluate a new and effective approach for predicting dyslipidemia in steel workers. Frontiers in Bioengineering and Biotechnology 8, 839. doi:10.3389/fbioe.2020.00839.

Xiang, Y., Ji, H., Zhou, Y., Li, F., Du, J., Rasmy, L., Wu, S., Zheng, W.J., Xu, H., Zhi, D., et al., 2020. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. Journal of medical Internet research 22, e16981. doi:10.2196/16981.

Yadav, P., Steinbach, M., Kumar, V., Simon, G., 2018. Mining electronic health records (ehrs) a survey. ACM Computing Surveys (CSUR) 50, 1–40. doi:10.1145/3127881.

Yang, F.C., Lee, A.J.T., Kuo, S.C., 2016. Mining health social media with sentiment analysis. Journal of medical systems 40, 1–8. doi:10.1007/s10916-016-0604-4.

Yang, Y., Guo, J., Ye, Q., Xia, Y., Yang, P., Ullah, A., Muhammad, K., 2021. A weighted multi-feature transfer learning framework for intelligent medical decision making. Applied Soft Computing 105, 107242. doi:10.1016/j.asoc.2021.107242.

Yang, Y., Hu, J., Liu, Y., Chen, X., 2020. Doctor recommendation based on an intuitionistic normal cloud model considering patient preferences. Cognitive computation 12, 460–478. doi:10.1007/s12559-018-9616-3.

Ye, C., Fu, T., Hao, S., Zhang, Y., Wang, O., Jin, B., Xia, M., Liu, M., Zhou, X., Wu, Q., et al., 2018. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. Journal of medical Internet research 20, e9268. doi:10.2196/jmir.9268.

Ye, Y., Zhao, Y., Shang, J., Zhang, L., 2019. A hybrid it framework for identifying high-quality physicians using big data analytics. International Journal of Information Management 47, 65–75. doi:https://doi.org/10.1016/j.ijinfomgt.2019.01.005.

Yoo, H., Han, S., Chung, K., 2020. A frequency pattern mining model based on deep neural network for real-time classification of heart conditions, in: Healthcare, MDPI. p. 234. doi:10.3390/healthcare8030234.

Yuan, H., Deng, W., 2021. Doctor recommendation on healthcare consultation platforms: an integrated framework of knowledge graph and deep learning. Internet Research 32, 454–476. doi:https://doi.org/10.1108/INTR-07-2020-0379.

Zhang, B., Ren, H., Huang, G., Cheng, Y., Hu, C., 2019. Predicting blood pressure from physiological index data using the svr algorithm. BMC bioinformatics 20, 1–15. doi:10.1186/s12859-019-2667-y.

Zhang, L., Fabbri, D., Lasko, T.A., Ehrenfeld, J.M., Wanderer, J.P., 2018. A system for automated determination of perioperative patient acuity. Journal of medical systems 42, 1–11. doi:10.1007/s10916-018-0977-7.

Zhang, O., Minku, L.L., Gonem, S., 2021. Detecting asthma exacerbations using daily home monitoring and machine learning. Journal of Asthma 58, 1518–1527. doi:10.1080/02770903.2020.1802746.

Zhang, Y., Sun, Y., Phillips, P., Liu, G., Zhou, X., Wang, S., 2016. A multilayer perceptron based smart pathological brain detection system by fractional fourier entropy. Journal of medical systems 40, 1–11. doi:10.1007/s10916-016-0525-2.

Zhao, Y., Fesharaki, N.J., Liu, H., Luo, J., 2018. Using data-driven sublanguage pattern mining to induce knowledge models: application in medical image reports knowledge representation. BMC medical informatics and decision making 18, 1–13. doi:10.1186/s12911-018-0645-3.