COLLNET Journal of Scientometrics and Information Management
ISSN : 0973 – 7766 (Print) 2168 – 930X (Online)
Vol. 15(2) December 2021, pp. 225–240
DOI : 10.1080/09737766.2021.1958659

Taylor & Francis
Taylor & Francis Group

Discovering the evolution of artificial intelligence in cancer research using dynamic topic modeling

Shahab Mosallaie [#]
Mahdi Rad [$]
Andrea Schiffauerova [@]
Concordia Institute for Information Systems Engineering (CIISE)
Concordia University
H3G 1M8
Montreal, QC
Canada

[#] shahab.mosallaie@mail.concordia.ca
[$] rad.mahdii@gmail.com
[@] andrea@encs.concordia.ca

Ashkan Ebadi *
Digital Technologies Research Centre
National Research Council Canada
H3T 2B2
Montreal, QC
Canada

and

Concordia Institute for Information Systems Engineering
Concordia University
H3G 1M8
Montreal, QC
Canada

ashkan.ebadi@nrc-cnrc.gc.ca

*(Corresponding Author)

Shahab Mosallaie
Mahdi Rad
Andrea Schiffauerova
Ashkan Ebadi

The rapid growth of healthcare data in recent years calls for more advanced and efficient analytic techniques. Artificial intelligence facilitates finding insightful patterns in massive high-dimensional data. Considering the latest movements towards using machine learning and deep learning techniques in the medical domain, in this study, we focused on the publications in which researchers employed artificial intelligence techniques for cancer diagnosis and treatment. Using dynamic topic modeling and natural language processing techniques, we analyzed the contents and trends of more than 12,000 scientific publications within the period of 2000 to 2018, extracted from two different sources, i.e., Elsevier's Scopus and PubMed. While drawing the landscape of cancer research, our results also shed light on the evolution of artificial intelligence techniques and algorithms used for cancer diagnosis and treatment. Our findings confirm that modern computer science algorithms are being widely applied to extract patterns from large-scale medical images to cure different types of cancer with a special focus on deep learning techniques in recent years.

Keywords: *Artificial intelligence, Cancer research evolution, Natural language processing, Dynamic topic modeling, Machine learning.*

1. Introduction

Cancer was identified as the second reason for death all over the world in 2018 with more than 9.5 million reported deaths [1]. Despite being a deadly disease, a large number of cancer patients can be cured if the disease is diagnosed in the early stages [1]. This highlights the importance of strategies and processes for cancer diagnosis to save many patients' lives who are suffering from this deadly disease [2]. Computer science algorithms embedded in decision support systems can help physicians diagnose cancer faster and more accurately, thereby improving cancer treatment and patients' outcomes [3].

A large volume of data is generated in different sectors of the healthcare industry [4]. Recent developments in information technology have led to an increase in the volume of digital data [5], the medical domain is no exception. In 2008, 13% of physicians reported that they have a basic electronic health record (EHR) system [6], while at the end of 2012, the number increased to 40% [7]. Due to the huge volume of data in healthcare, the conventional analytic methods seem to be not very efficient anymore [4]. Hence, researchers have been employing more advanced approaches like artificial intelligence (AI), machine learning, and deep learning to process, analyze, and interpret massive data [3], [8], [9].

AI techniques are applied to various types of healthcare data, from structured to unstructured, mainly focusing on cancer, neurology, and cardiology as major disease areas [9]. Deep learning techniques are applied to medical imaging and they have shown promising results for disease detection. For example, recent studies showed valuable use of AI and deep learning in particular in the diagnostics and detection of lung cancer [10], [11], breast cancer [12], pancreatic cancer [13], prostate cancer [14] or cervical cancer [15]. Some recent studies, e.g., [16]–[19], overview latest advances and various applications, and provide recommendations for the use of deep learning in cancer diagnostics.

Research publications are one of the major channels for sharing scientific knowledge and discoveries. There are several publication databases in the medical field with the most comprehensive and common ones to be PubMed, Scopus, and Web of Science [20], [21]. PubMed is a search engine that facilitates searching across several literature resources including MEDLINE as the largest component of PubMed, PubMed Central (PMC), and Bookshelf which covers an archive of books, reports, and databases in the field of biomedical and health [22]. Scopus and Web of Science are abstract and citation databases covering about 24,000 titles and 9,200 journals, respectively. Falagas et al. [20] investigated differences between the aforementioned databases and found Scopus to be the most comprehensive database, covering more journals and articles. On the other hand, PubMed is an optimal tool in medical research covering approximately 30,000 journals [23], granting researchers access to the biomedical research articles free of charge while being updated routinely and quickly [24]. To obtain a comprehensive view of the application of AI in cancer diagnosis/treatment, it is suggested in the literature to consider multiple data sources [24].

In recent years, advanced computer science techniques such as natural language processing (NLP) and topic modeling have been being applied to massive text datasets in

many diverse domains such as bioinformatics [25], social media [26], behavior modeling in crowded traffics [27], transportation [28], management [29], federally funded research alignment [30], communication [31], marketing [32], smart factory [33], and COVID-19 research [34], to discover latent research themes and analyze their trends.

The number of publications using AI to solve medical problems has tripled from 2016 to 2019, with the highest interest in cancer research [35]. These publications that cover a broad range of techniques being applied to different types of medical data have led to significant scientific developments in cancer diagnosis and treatment [36]. Deep learning models, in particular, have showed promising performance in detecting various types of cancer such as lung and breast cancers, mainly using medical imaging [37]–[39]. Tariq et al. [40] did a comprehensive review over automated breast cancer detection techniques using multiple imaging modalities and found deep learning techniques as efficient models for the purpose when they are provided with huge training datasets.

One of the applications of topic modeling is helping researchers to survey the research trends in the related scientific literature [41]. There exist some studies that used advanced text mining techniques to mine medical literature. For example, Geletta et al. [42] investigated 252,847 studies from the Clinical Trials Transformation Initiative (CTTI) and used the Latent Dirichlet Allocation (LDA) topic modeling technique to predict the completion or termination of trial studies. In another study, Tran et al. [41] investigated abstract of publications from 1991 to 2018 and used LDA topic modeling to analyze the trend of AI in cancer care research. Their study only provides a static view of the research landscape since they did not take the temporal aspect into the account.

In this study, we used multiple data sources to analyze the content of research publications about the application of AI in cancer diagnosis and treatment within 2000-2018, providing a concrete landscape of the state of the research as well as its evolution. The objectives of this study are: 1) to dynamically quantify the development of research in a highly multidisciplinary ecosystem, 2) employ multiple data sources to be able to compare the patterns and obtain a better picture of the state of research, and 3) specifically, focus on the developments in the field in the current century. We used NLP and dynamic topic modeling (DTM) to extract hidden patterns and latent research themes, and to analyze their evolution. To the best of our knowledge, this is the first study that uses multiple data sources and considers the temporal aspect in analyzing AI in cancer publications. The rest of the paper proceeds with the "Materials and Methods" section which describes the data and techniques in greater detail, the findings of the research are presented in the "Results" section and are discussed in the "Discussion" section. The paper concludes in the "Conclusion" section.

2. Materials and Methods

Sample and data

We used the following search query to extract AI in cancer research publications from Elsevier's Scopus and PubMed within the [2000, 2018] time interval: ("artificial intel-

ligence" OR "machine learning" OR "deep learning" OR "neural network" OR "neural net") AND ("cancer"). We only included journal articles, conference papers, book chapters, and books. This resulted in 10,071 and 2,206 publications collected from Scopus and PubMed, respectively.

Variables of interest

Date of publication, title, and abstract of the publications are the three most important variables in this analysis. Date of the publication is used in the model to capture the temporal evolution of the latent themes. Titles and abstracts provide the textual input from which the latent themes are extracted. We decided to concatenate titles and abstracts as the abstract provides a condensed representation of articles with detailed information in comparison with publications' titles, but titles can also contain some complementary information such as specific keywords or key phrases about the research [30].

Data analysis

We considered three data scenarios for the analysis: 1) Scopus publications only, 2) PubMed publications only, and 3) the intersection of Scopus and PubMed publications. We coded the entire analytics pipeline in Python. Figure 1 shows the conceptual flow of the analytics pipeline. We merged titles and abstracts of the collected publications, preprocessed the textual data, and used the DTM algorithm [43] in each of the three data scenarios

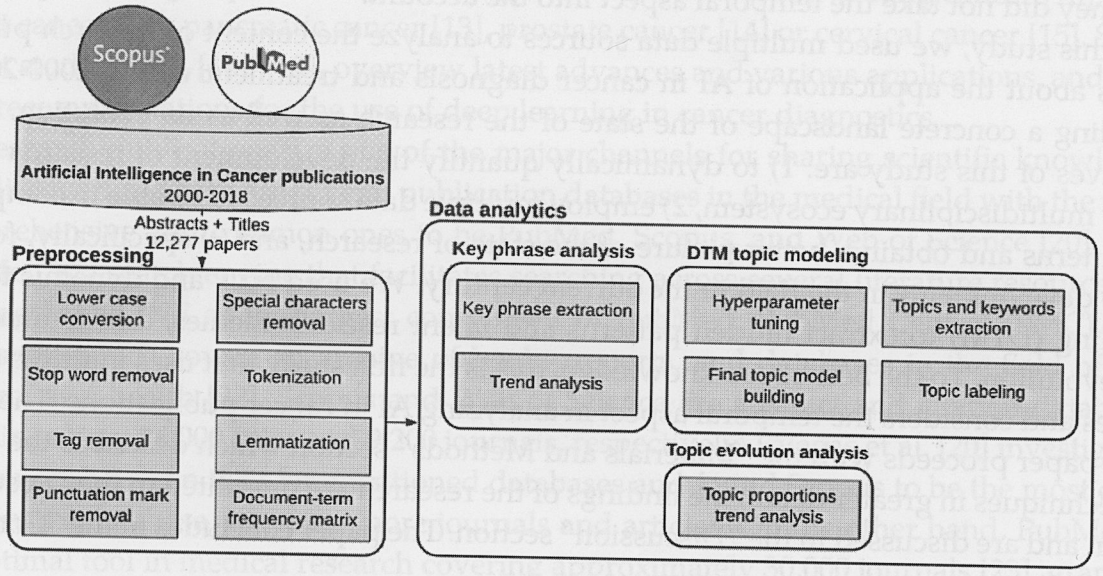

Figure 1

The conceptual flow of the analytics pipeline. The pipeline contains three main components, i.e., data collection, processing, and data analytics. In the data collection step, AI in cancer publications for the period of 2000-2018 are extracted. Then, abstracts and titles of the publications are merged and preprocessed. Next, key phrases are extracted, their trends over time are analyzed, and the DTM model is built to extract latent research topics. Finally, the trends and evolution of the extracted research topics are investigated.

to extract latent research themes out of the target publications. We did several preprocessing steps including but not limited to lowercase conversion, stop words and punctuation removal, tokenization, and lemmatization. We considered uni- and bi-grams to extract key phrases and investigated their trends over the examined period. We then used uni-grams to calculate the document-term frequency (TF) matrix that the DTM algorithm consumed to build the model. We performed intensive hyperparameter tuning on the DTM model. The DTM algorithm requires the number of topics to be defined in advance. We first built several LDA [44] baseline models for each dataset to find a range for the number of topics by investigating the coherence of the extracted topics. We found the best range for the number of topics to be in [5, 8]. Next, we built four separate DTM models for each dataset, with five to eight topics, and assessed the quality of their topics both quantitatively with *CV* coherence scores [45], and qualitatively, by verifying with three domain experts. The DTM models with six topics were found to be the best for Scopus and PubMed. For the Scopus-PubMed intersection data scenario, the optimal number of topics was found to be five. In addition to the number of topics, we further tuned other hyperparameters such as chunk size.

The DTM algorithm does not assign a representative label to the extracted topics. To reduce the subjectivity impact, we used a panel of three experts to analyze the expanded set of keywords generated for each topic and to assign a short informative label to them. Since in this study we were interested in finding the main "AI in cancer" research themes, the experts were tasked to find more generic and inclusive topic labels, preferably with a simple and interpretable name rather than a phrase. The extracted topics for the three data scenarios are as follows:

- **Scopus only:** 1) Medical image analytics, 2) Cancer survival, 3) Drug design, 4) Machine learning, 5) Cancer genomics, and 6) Clinical decision support systems
- **PubMed only:** 1) Medical image analytics, 2) Cancer survival, 3) Drug design, 4) Machine learning, 5) Cancer genomics, and 6) Colorectal cancer
- **Scopus-PubMed intersection:** 1) Medical image analytics, 2) Cancer survival, 3) Drug design, 4) Machine learning, and 5) Clinical decision support systems

The extracted topics are almost the same in the three data scenarios. The only difference between Scopus only and PubMed only scenarios is that the "clinical decision support system" is observed in the former while "colorectal cancer" is instead observed in the latter.

3. Results

The total number of publications in Scopus and PubMed datasets is 12,277 (Scopus dataset with 10,071 and PubMed dataset with 2,206 publications). The two datasets overlap significantly with ~86% of PubMed publications being covered in Scopus as well. The overlap rate ranges from ~81% in 2000 to ~76% (i.e., the minimum) in 2018, with a ~95%

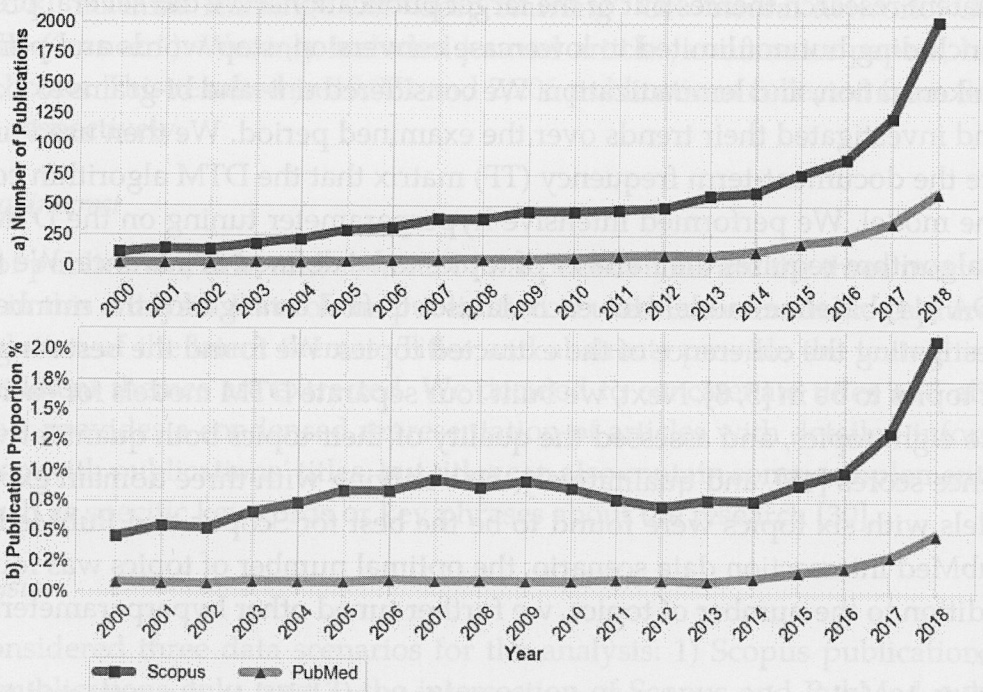

Figure 2

AI in cancer publications trend in PubMed and Scopus, a) exact number of AI in cancer publications, b) proportion of AI in cancer publications over all cancer research publications.

peak in 2006. There are 331 (3.2%) and 8,196 (78.8%) unique publications in PubMed and Scopus datasets, respectively.

Figure 2-a shows the number of publications in Scopus and PubMed. As seen, the number of publications has increased continuously from 2000 to 2018 in both datasets. There has been a significant increase in the final two years such that the number of publications in Scopus had a ~40% and ~65% growth in 2017 and 2018, respectively. In PubMed, a growth rate of 67% is observed in 2018. Although this could be due to the better coverage of publications in the final years, it could also partially imply the higher attention of the researchers to the subject topic, i.e., AI in cancer. Figure 2-b depicts the proportion of AI in cancer publications over all cancer research publications. From the figure, an almost steady trend is observed for PubMed publications till 2013, being followed by an increasing trend afterwards. This recent growth might reflect the traction of more researchers. Although the trend for Scopus is slightly different in the beginning periods, a sharp increase is observed after 2016.

To initially analyze the evolution of computer science algorithms, we extracted the most frequent computer science algorithms/techniques mentioned in the collected publications, i.e., machine learning, random forests, logistic regression, principal component analysis (PCA), deep learning, and convolutional neural network (CNN), and analyzed their proportion over time in Scopus and PubMed (Figure 3). As seen, in both Scopus and PubMed datasets, "machine learning" has the highest proportion with some fluctuations

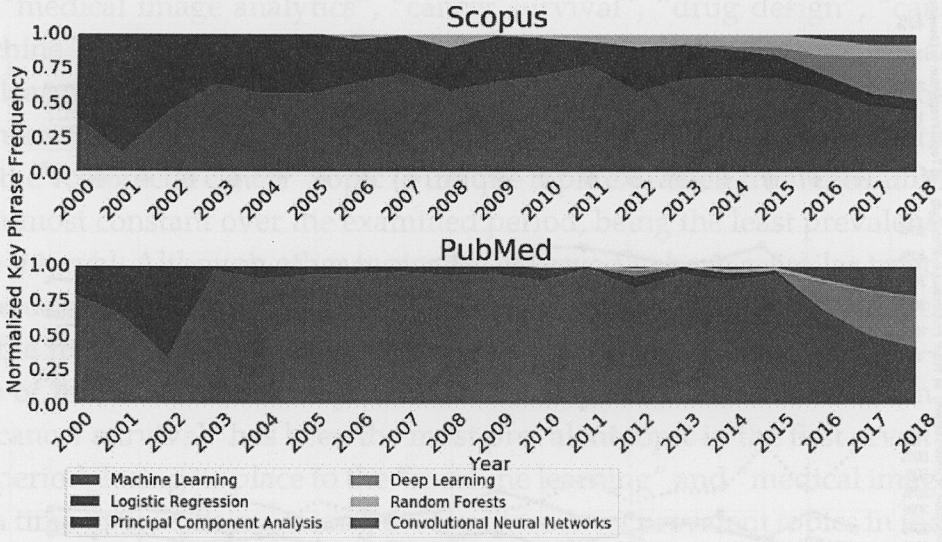

Figure 3

The temporal trend of computer science algorithms used in AI in cancer research represented by the proportions of top algorithmic key phrases observed in Scopus and PubMed.

from 2000 to 2002, especially in PubMed, and almost a constant trend from 2003 to 2015. The "machine learning" key phrase follows a decreasing trend after 2015 in both datasets. In the Scopus dataset, the "random forests" algorithm first appeared among the most frequent key phrases in 2005 and its proportion slightly increases after 2009, whereas in PubMed it is first seen in 2009, following an almost constant trend until the final period. Additionally, in the Scopus dataset, "deep learning" and "convolutional neural network" are observed for the first time in 2012 and 2015, respectively, and their proportion increases thereafter. Similarly, in the PubMed dataset, we can see the appearance of "deep learning" and "convolutional neural network" key phrases in 2011 being followed with a sharp increase after 2015. In both datasets, the trends for "logistic regression" has decreased over time. These observations may partially confirm that researchers active in the healthcare and cancer diagnosis domain have gradually shifted from conventional statistical analysis to more advanced computer science algorithms due to several factors such as conventional techniques limitations in handling massive data and/or new types of data, e.g., medical images [46].

Topics prevalence in Scopus and PubMed datasets as well as their intersection is depicted in Figure 4. As explained in section 2, six topics were extracted from the Scopus dataset. As seen in Figure 4-a, the "machine learning" topic has been the most prevalent topic in almost all the examined period, being replaced by the "medical image analytics" topic in the last three years. This would imply the increasing interest of researchers in using machine learning algorithms to solve complex problems such as cancer diagnosis. This finding is in line with Kourou et al. [47]. Interestingly, after 2015, the "medical image analytics", being represented by deep neural networks key phrases, e.g., CNNs, takes the place of the "machine learning" topic as the most prevalent topic. This confirms our

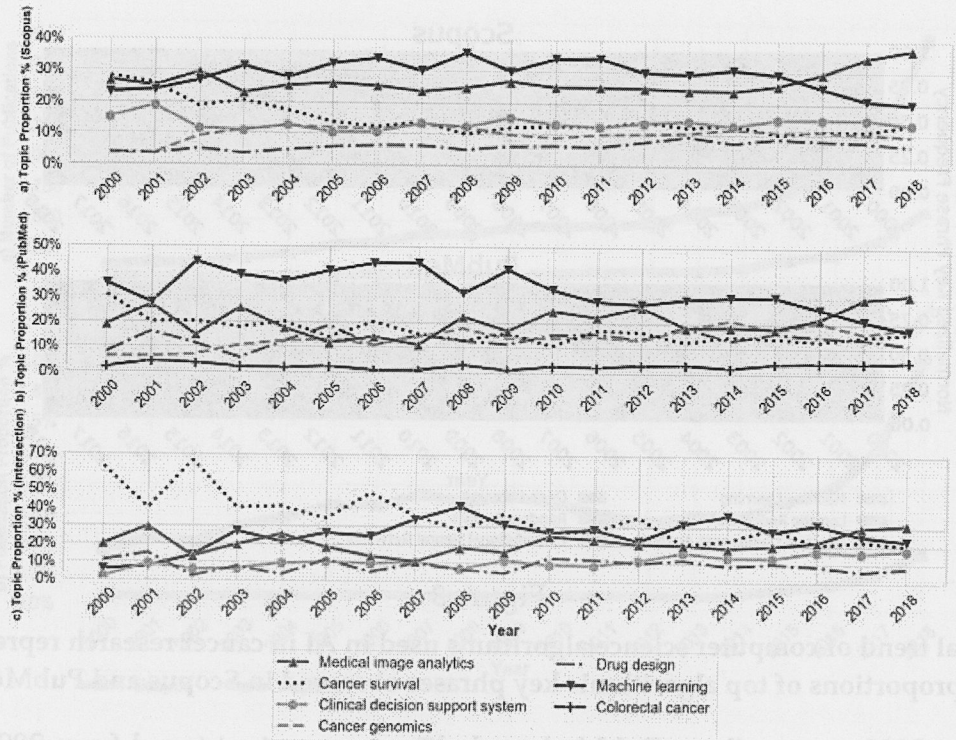

Figure 4

Topic prevalence from 2000 to 2018 in a) Scopus, b) PubMed, c) the intersection of Scopus and PubMed.

findings in the previous section about the application of deep learning approaches to analyze new data types such as medical images (Figure 3). Medical image analytics has been used in practice for many years especially in computer-aided diagnosis (CAD) systems to assist radiologists and clinicians [48]. A large number of publications have investigated the application of CNN algorithms for diagnosing different types of cancers such as breast cancer, lung cancer, and prostate cancer [49]–[53]. The "cancer survival" topic proportion is high at the beginning of the examined period, having an almost constant prevalence after 2008, despite some fluctuations. This might also indicate the application of AI in medical subfields. Our data also suggests that the "clinical decision support system" and "cancer genomics" research themes have attracted the attention of researchers, following an almost steady trend after 2002. This results from the rapid growth of healthcare data in different forms including genomic datasets and the need for intelligent decision support systems to analyze and process the massive amount of data quickly and efficiently [54], [55]. The "drug design" topic's proportion has slightly increased from 2000 to 2013 and it remained almost constant afterward. The appearance of this topic was expected as researchers are using AI and advanced techniques, for example, in the drug development process, such as drug repurposing, predicting the mode-of-action of compounds, selection of a population for clinical trials, etc., where AI contributed to a higher efficiency and lower research and development (R&D) costs [56]. Figure 4-b shows the prevalence of the topics for the PubMed dataset. Six topics were extracted from the PubMed dataset

including "medical image analytics", "cancer survival", "drug design", "cancer genomics", "machine learning" and "colorectal cancer". Comparing Figures 4-a and 4-b, the "machine learning" and "medical image analytics" topics have been the most prevalent research themes in both Scopus and PubMed datasets, having similar patterns. The proportion of the "colorectal cancer" topic (a unique topic extracted from the PubMed dataset) remained almost constant over the examined period, being the least prevalent topic in the entire time interval. Although other topics are observed to have a similar trend to the ones seen in Scopus, fewer fluctuations are observed in the Scopus topics prevalence that might be due to the higher number of publications in the Scopus dataset. Figure 4-c shows the prevalence of the five extracted topics from the Scopus-PubMed intersection dataset. As seen, the "cancer survival" has been the most prevalent topic in the first seven years of the examined period, losing its place to the "machine learning" and "medical image analytics" topics from time to time. Being among the top three most prevalent topics in the entire time interval, the "medical image analytics" becomes the most prevalent topic in the final year.

4. Discussion

As an aggressive disease with a low median survival rate, cancer has a long and costly treatment process. Early diagnosis can enhance patients' survival chances. Over the years, scientists have been using statistics and computational methods to diagnose/predict the disease. With the emergence of AI and learning techniques, many scientists are applying machine/deep learning to clinical cancer research where the performance of the cancer prediction models has been promising [57]. For example, deep convolutional neural networks were shown to improve the diagnostic accuracy of solid tumors in thyroid cancer [58], or classification of malignant and benign masses in digital breast tomosynthesis [59]. Apart from the performance, such technologies are providing unique opportunities to analyze new data types such as medical imaging. For example, AI has been applied to medical imaging such as magnetic resonance imagery (MRI) and computerized tomography (CT) scans, facilitating the analysis of new data types.

By doing a cross publication search engine study within the period of 2000-2018 and using dynamic topic modeling and natural language processing, this study provides a concrete perspective on how scientists are employing AI for cancer detection and treatment, highlighting the current research trends. Our findings suggest a shift from conventional analytic techniques towards learning techniques. Besides, medical image analytics was found to be a prevalent theme that has been increasingly attracting researchers' attention over time. Deep learning techniques, specifically, are assisting data scientists to analyze and interpret imaging data more precisely [60], [61], ensuring fewer false positives than radiologists [57]. Although our results show that the importance of AI is being increasingly recognized in the medical domain, a tight collaboration between computer scientists and medical experts would play a key role in ensuring the continued success of this interdisciplinary research.

The complexity of medical chemistry research calls for the application of emerging technologies in the design of new drugs [62]. Machine learning and AI are being used in the new drug discovery process, aiming to make it faster, cheaper, and more effective [63]. We identified "drug design" as one of the main research themes. Considering that data for this type of research is being available more than ever, and the number of academic and industrial labs employing AI for drug design is increasing [64], we expect to observe this topic as one of the main research themes in the coming years as well, even with a continuous growth. As a deadly disease affecting more than 18 million people worldwide annually, cancer is a disease of the genes being caused by mutations in genomes [65]. The "cancer genomics" topic was observed to be one of the most prevalent topics, especially in the PubMed dataset in the final years. Hence, the results suggest that AI is assisting researchers to perform a more concrete analysis on large quantities of genome data which may result in a better understanding of the disease while adapting the treatment to the molecular characteristics of each patient [65].

The incorporation and increasing availability of clinical decision support systems (CDSS) in healthcare settings is major progress that supports clinicians with the decision-making process that would lead to improved quality of care while minimizing the costs [66]. The CDSS topic was identified as one of the most prevalent topics in our study. Considering the potential of AI for various types of cancer diagnosis, we expect that AI-powered CDSS research will continue to grow resulting in a paradigm shift in cancer diagnosis and treatment, as AI is foreseen to keep helping scientists to overcome the challenges of cancer diagnosis [57]. This development could be facilitated by the increased availability of digital medical data as well as the growth of medical data scientists. Finally, the appearance of "colorectal cancer" among the topics extracted from the PubMed dataset explains the importance of this type of cancer in the eyes of the researchers. According to a recent study, colorectal cancer is the second leading cause of cancer death and the third most common cancer in men and women [67]. Hence, special attention to this field of research and continuous support would be encouraged.

5. Conclusion

The rapid growth of digital data and recent advancement in modern analytics techniques have opened a new gate of opportunities to analyze massive datasets [68] and new data types such as text data. Natural language processing and topic modeling offer an objective and quantitative approach to measure latent patterns. We proposed an approach to better understand the "AI in cancer" research landscape and investigated how researchers contributed to this innovation ecosystem and used advanced techniques to improve patients' outcomes by extracting the main research themes and assessing their temporal evolution. The DTM algorithm allowed us to dynamically calculate topic proportions over time. Our findings confirm the growth of using machine/deep learning techniques in analyzing healthcare data. Especially, medical image analytics and deep convolutional neural networks were found to be promising directions in recent years in analyzing enormous