Li *et al. Journal of Big Data* (2024) 11:61
https://doi.org/10.1186/s40537-024-00929-2

Journal of Big Data

RESEARCH Open Access

High-performance computing in healthcare:an automatic literature analysis perspective

Jieyi Li[1*], Shuai Wang[1], Stevan Rudinac[1] and Anwar Osseyran[1]

*Correspondence:
j.li3@uva.nl

[1] Amsterdam Business School, University of Amsterdam, Plantage Muidergracht 12, Amsterdam 1018 TV, The Netherlands

Abstract

The adoption of high-performance computing (HPC) in healthcare has gained significant attention in recent years, driving advancements in medical research and clinical practice. Exploring the literature on HPC implementation in healthcare is valuable for decision-makers as it provides insights into potential areas for further investigation and investment. However, manually analyzing the vast number of scholarly articles is a challenging and time-consuming task. Fortunately, topic modeling techniques offer the capacity to process extensive volumes of scientific literature, identifying key trends within the field. This paper presents an automatic literature analysis framework based on a state-of-art vector-based topic modeling algorithm with multiple embedding techniques, unveiling the research trends surrounding HPC utilization in healthcare. The proposed pipeline consists of four phases: paper extraction, data preprocessing, topic modeling and outlier detection, followed by visualization. It enables the automatic extraction of meaningful topics, exploration of their interrelationships, and identification of emerging research directions in an intuitive manner. The findings highlight the transition of HPC adoption in healthcare from traditional numerical simulation and surgical visualization to emerging topics such as drug discovery, AI-driven medical image analysis, and genomic analysis, as well as correlations and interdisciplinary connections among application domains.

Keywords: Literature analysis, Topic models, High performance computing, Healthcare, Research trends

Introduction

High Performance Computing (HPC) is increasingly becoming an indispensable resource in healthcare research due to its capabilities in addressing complex and data-intensive tasks [1–3]. The exponential growth of health data next to simulation and modeling drives the adoption of HPC, which encompasses i.a., genomic sequencing, biomedical imaging, electronic health records (EHRs), and wearable device data [4–8]. Effectively managing and analyzing such data poses significant challenges in storage, management, and analysis, necessitating the computational power offered by HPC. In genomics, HPC allows researchers to analyze genomic data at a scale and speed previously impossible, revealing genetic bases of various diseases and helping develop

personalized treatments [9, 10]. Similarly, HPC plays a crucial role in drug discovery and computational modeling [11, 12]. Drug discovery processes are traditionally time and resource-intensive. However, HPC allows researchers to conduct molecular dynamics simulations to understand drug-protein interactions at the atomic level, speeding up the process and reducing costs [13–15]. Furthermore, HPC-enabled computational modeling allows the simulation of biological systems or disease progression, providing insights to inform treatment strategies [16, 17]. The deployment of HPC in Artificial Intelligence (AI) for healthcare has also experienced substantial growth. HPC enables the implementation of advanced AI techniques, which necessitate substantial computational power to efficiently handle extensive data volumes and complex deep neural networks [18, 19]. Biomedical imaging stands as a prime example where HPC assumes a crucial role [20]. By leveraging HPC, AI tools can swiftly process and analyze high-resolution images, facilitating real-time analysis of complex imaging data and leading to expedited and more precise diagnoses [21, 22]. In addition, HPC-enabled convergence of AI and simulation has significantly improved the quality and speed of traditional simulation in healthcare [23, 24].

Investigating the literature on HPC adoption in healthcare can generate valuable insights that are beneficial for the business and economic side of healthcare by providing a comprehensive understanding of the current landscape and potential future directions. These insights can guide strategic planning and investment decisions of HPC in healthcare businesses, highlighting promising areas for further exploration and development. However, the substantial volume of literature, combined with the rapid pace of technological advancements, makes manual analysis very difficult and time consuming. As a result, there is a recognized need for an automated literature analysis framework to accurately process the vast corpus of literature, transforming it into meaningful insights for business or strategic decision making within the healthcare sectors.

Topic modeling, a family of typically unsupervised machine learning approaches, aims at discovering hidden semantic structures, or 'topics', within a corpus of text [25]. The underlying principle of topic modeling is to classify text documents into different topics based on the frequency and co-occurrence of words [26]. The strength of topic modeling lies in its capacity to handle large and unstructured datasets, rendering it an invaluable tool for exploratory data analysis. Prominent techniques employed in topic modeling include Latent Dirichlet Allocation (LDA) [27], Non-negative Matrix Factorization (NMF) [28], and Latent Semantic Analysis (LSA) [29]. With a broad range of applications in areas such as text mining, information retrieval, and digital humanities, topic modeling continues to garner considerable interest [30–32]. Topic modeling has become an increasingly popular tool in scientific research and literature review [33–35]. Its usage spans various scientific research fields, including marketing, medical, and social sciences [34, 36, 37]. In the context of literature reviews, topic modeling has been used to identify trends and patterns in large bodies of literature. For instance, it has been used to analyze collective behavior and social movements by sociologists [37], and also been adopted to understand the big data themes from biomedical research [38].

LDA is arguably one of the most widely used algorithms for topic modeling. However, it has been characterized by certain constraints, including the prerequisite of data cleaning and pre-processing, the requirement to specify model parameters such as α

(document-topic density), β (topic-word density) and topic numbers hyperparameters. Moreover, the challenges associated with the interpretability and validation of the generated topics also need to be addressed [39, 40], which is why alternatives such as entity linking (EL) have been frequently deployed, especially in case of shorter texts [41]. In response to these limitations, recently developed deep learning algorithms such as Top2Vec, offer alternative approaches for topic modeling [40]. Top2Vec transforms each word in a text collection into a vector representation within a semantic space using an encoding model such as doc2vec or state of the art transformers. Consequently, it autonomously identifies topics within the text and generates jointly embedded topic, document, and word vectors. Comparative studies between LDA and Top2Vec have been conducted, revealing that Top2Vec tends to yield qualitatively superior results compared to LDA [42, 43].

In this study, we propose an automatic literature analysis framework, using a complex question of the impact of HPC on healthcare as the test-bed. The primary contributions of this work are twofold:

1. We present an automatic literature analysis framework, from document (i.e. scientific article) retrieval and analysis to various interactive visualizations to depict research trends, topics evolution, interconnection across research areas and high impact papers. This adaptable pipeline can be easily applied to other domains by modifying the initial query keywords.
2. By deploying the automated literature analysis framework, we investigate the research trends of HPC utilization in healthcare. Our analysis reveals notable shifts in research focus, spanning from visualization and rendering in surgical practice and traditional numerical simulation to emerging topics such as drug discovery, AI-driven medical image analysis, and genomic analysis. These insights provide valuable indications for future investment and strategic development of HPC within the healthcare sector, guiding decision-making and resource allocation.

Materials and methods

In this section, we provide a comprehensive description of the data and analysis process employed in our study. Figure 1 illustrates the automatic literature review framework implemented in our study, which consists of four distinct stages: paper retrieval and extraction, data preprocessing, topic modeling, and visualization. The subsequent sections provide detailed explanations of each stage. The code for the automated literature review framework is publicly available in a GitHub repository.[1]

Paper retrieval and extraction

Data source

Scopus[2] has been utilized in our study as it is the largest publicly available abstract and citation database of peer-reviewed literature, including scientific journals, books and conference proceedings.

[1] https://github.com/tuohai1992/Automatic-literature-analysis-pipeline..

[2] https://www.elsevier.com/solutions/scopus.

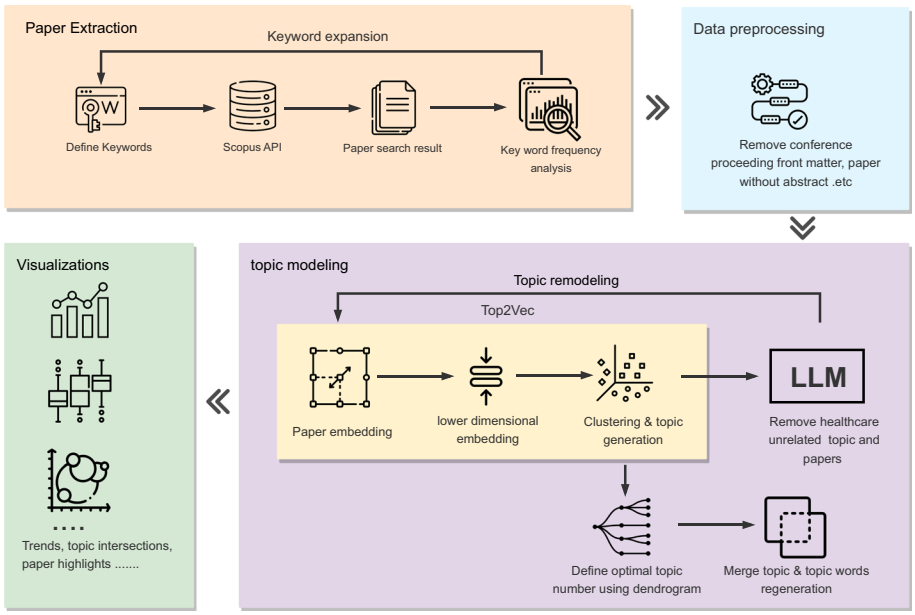

Fig. 1 Automatic literature review pipeline consists of four stages: paper retrieval and extraction, data preprocessing, topic modeling, and visualization

Type of articles included in the study

In this study, conference papers, journal papers, books or book chapters, editorial materials, early access workshop papers and presentations published in English are included. Conference proceeding front matters, meeting abstracts, errata and papers without abstracts are excluded from the analysis. Since the publication year 2023 is not yet complete, studying trends beyond 2022 may introduce bias. Therefore, we limit our analysis to publications up until end of 2022.

Paper retrieval pipeline

We establish an automatic paper record retrieval and query expansion pipeline through the Scopus API.[3] During the initial phase of query construction, we take into account the co-occurrence of HPC keywords such as 'high performance computing' and 'supercomputing' along with the most prevalent health-related terms like 'healthcare', 'health', 'clinic', 'medicine', 'disease', and 'treatment'. With each iteration of the literature search, we implement query expansion by incorporating additional relevant keywords extracted from the retrieved scientific paper documents. To achieve this, we perform a frequency analysis of keywords extracted from the titles and abstracts. Keywords that are identified as HPC or healthcare-related and occur with a frequency of 50 or higher, but are not previously included, are considered as new keywords and subsequently added to the expanded query. Such automatic query expansion process results in inclusion of the synonyms of HPC, such as 'high-performance computing', 'high performance computer', and 'supercomputer', along with broader or narrower healthcare-related terms like 'patient', 'diagnosis', 'drug', 'therapy', 'pharmaceutical', 'surgery', and others. This process is

[3] https://dev.elsevier.com/.

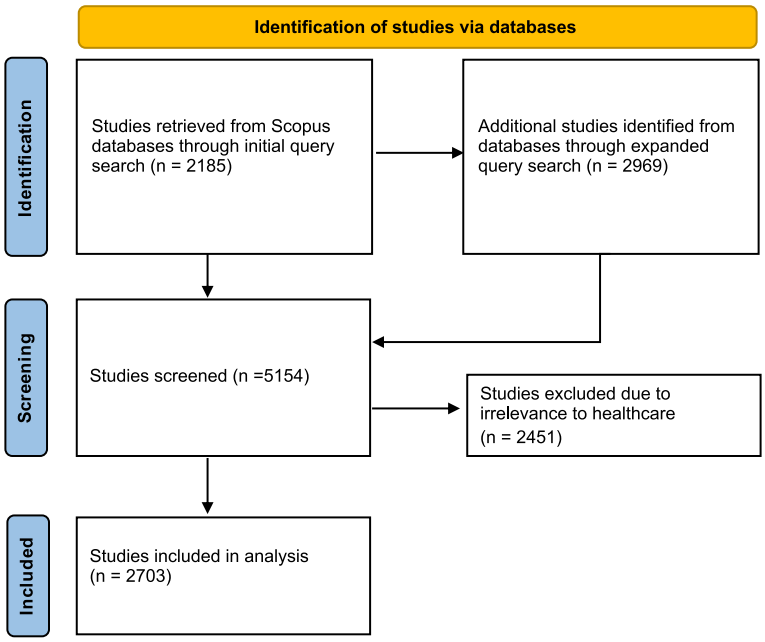

Fig. 2 Flow diagram of our automatic literature analysis

repeated until no new keywords emerge, indicating the completion of the search query expansion. Fig. 2 presents the flowchart for the studies included in our automated literature analysis. Initially, 2185 studies were identified through an initial query search, and an additional 2969 studies were included via query expansion. After excluding studies not directly related to healthcare ("Identify outlier topics not focusing on HPC adoptionin healthcare and remodeling" for detailed explanations), a total of 2703 studies were incorporated into the analysis. The extracted literature serves as an input for further processing in the subsequent steps.

Data preprocessing

Based on the output obtained from Scopus, we retain the title, abstract, publication date, and citation number for subsequent topic modeling and visualization. Given that the Top2Vec algorithm does not require stop-word lists nor stemming, or lemmatization, the title and abstract of each literature piece are merged as the model input without further preprocessing.

Topic modeling

Algorithm choice

We adopt the Top2Vec model for topic modeling, an innovative unsupervised machine learning algorithm designed for automatic topic detection and document embedding [40]. This model is unique as it combines the strengths of word embeddings, dimensionality reduction, and density-based clustering to identify topics from a given set of documents without any prior knowledge or human intervention.

The first step in the Top2Vec algorithm involves transforming the documents into dense vector representations using chosen embedding algorithm to capture the semantic

meanings of the documents, including the context in which words are used, and represents them as high-dimensional vectors. This process results in a document embedding space where semantically similar documents are located close to each other. We process the input literature by employing five distinct embedding techniques available in the Top2Vec algorithm to generate combined document and word vectors. These techniques include doc2vec [44], two Universal Sentence Encoder models [45, 46], and two BERT models [47, 48]. We implement the CV coherence score in our topic modeling evaluation, a metric initially introduced by Röder et al. in their comprehensive examination of coherence measures for topic modeling algorithms [49]. The CV coherence score combines cosine similarity with Normalized Pointwise Mutual Information (NPMI). This selection is premised on the strong correlation that the CV coherence score maintains with human ratings, outperforming other evaluative measures. Consequently, we choose the embedding model demonstrating the highest CV coherence score for our study.

Once the documents are represented as vectors, the Top2Vec model applies the UMAP (Uniform Manifold Approximation and Projection) algorithm to reduce the dimensionality of these vectors [50]. UMAP is a manifold learning technique used for dimension reduction. It preserves both the local and global structure of the data, meaning that it maintains the distances between nearby points (local structure) and distant points (global structure). This results in a lower-dimensional space where clusters of document vectors represent unique topics.

Following the dimensionality reduction, the model uses HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a density-based clustering algorithm, to identify these clusters [51]. HDBSCAN works by finding regions of the reduced space where there are higher densities of document vectors, and it groups these together as clusters. Each cluster represents a unique topic in the document set and the centroid of each cluster is identified as a 'topic vector'. This topic vector is a point in the reduced space that best represents the semantic meaning of each topic. The topic vectors are subsequently converted back into the word space to provide interpretable topics. This is done by identifying the n-closest word vectors to the resultant topic vector. For each topic, the top 50 words are returned, arranged in order of proximity to the topic. One of the key advantages of the Top2Vec model is that it automatically determines the optimal number of topics based on the data. Additionally, it provides both the keywords for each topic and the documents that are most semantically similar to each topic, offering a comprehensive understanding of the topics present in the document set.

Identify outlier topics not focusing on HPC adoption in healthcare and remodeling

During the literature extraction process, articles that contain healthcare and high-performance related keywords within the title, abstract, or keyword sections are extracted via the Scopus API. However, initial topic modeling results reveal that certain articles do not relate to HPC adoption in healthcare. This discrepancy can be attributed to multiple factors. Firstly, while some articles include the terms 'high performance computing' and 'health' or 'healthcare' in their titles or abstracts, they predominantly address aspects of system health such as fault tolerance, job scheduling, and interconnection, rather than human health. This results in these articles being categorized under topics such as system architecture or networks. Secondly, many abstracts begin with a broad

statement, such as "high-performance computing has been widely used in several industries, including healthcare..." but the remainder of the abstract primarily discusses the adoption of HPC in domains other than healthcare.

The GPT-3 series model has demonstrated promising results in binary semantic text classification [52]. To fully automate our analysis pipeline and minimize potential bias from human judgment, we employed `text-davinci-003`, the most advanced model in the GPT-3 series, for non-healthcare related topic detection. Similar to studies [53, 54], following the first round of topic modeling that generates keywords, we design prompts using `the text-davinci-003` text completion API[4]. These prompts incorporate the top 20 keywords from each topic, aiming to produce binary outputs that determine whether the keyword combination for each topic pertains to the healthcare domain. To align with our expectation of binary output (`Yes` or `No`), we adjust the `max_tokens` parameter indicating the upper limit of tokens to be generated, to a smaller value of 5. The `temperature` parameter is set to 0 to limit the randomness of the generated responses, ensuring a more focused and deterministic output. All other hyperparameters are retained at their default values. If a topic is classified as unrelated to healthcare, the articles under that topic are identified as outliers and removed from the dataset. Once all topic keywords have been examined, the remaining literature proceed to the second round of topic modeling.

Identify optimal topic number using dendrogram

Upon evaluating the preliminary output from the Top2Vec model after the outlier detection phase, we observe that certain topics demonstrate substantial similarity and an increased granularity. Examples include multiple themes associated with virus and epidemic research, genomic research, and drug discovery (cf. Section and Fig. 3 for detailed observations). Merging similar topics to identify an optimal topic number could prove beneficial in subsequent analysis, providing a more overarching perspective on HPC adoption trends in healthcare.

Agglomerative hierarchical clustering with dendrogram is a technique used to aggregate similar data points or objects into clusters based on their pairwise distances [55]. This technique initiates with individual data points and sequentially merges these based on their proximity relationships. For the computation of similarities between topic vectors, cosine similarity has been utilized. This measure proves to be particularly beneficial when processing high-dimensional data, such as semantic word embeddings [56, 57]. Cosine similarity takes into account both the magnitude and direction of each vector, property making it more robust compared to the common alternatives like Euclidean distance, which considers only the magnitude. In the context of our hierarchical clustering, we have adopted the average linkage method [55].

The outcome is a dendrogram that displays hierarchical relationships. The methodology starts with the calculation of a distance matrix that captures the distances between topic vectors. The closest clusters are sequentially merged, and the distances are updated correspondingly. The dendrogram is constructed to portray the merging process, with

[4] https://platform.openai.com/docs/models/gpt-3-5.

branch lengths representing dissimilarity. By studying the dendrogram, an appropriate cutting point can be identified to determine the number of clusters. Hierarchical clustering provides a comprehensive visualization and organization of topics, aiding in understanding its inherent structure and interrelationships. Therefore, dendrogram has been utilized to visualize the topic merging process and to determine the optimal topic number.

Topic merging

Having determined the optimal number of topics using the dendrogram, we proceed to consolidate the topics. Top2Vec provides a function `hierarchical_topic_reduction` designed to decrease the number of topics identified by the algorithm[5]. However, the process operates by iteratively merging the smallest topic based on the number of associated documents with the most similar topic until the predetermined number of topics is reached. We conjecture that this approach, which prioritizes the size of the topics rather than their similarity for merging, may not be optimal. Specifically, the topic associated with the smallest cluster of documents in each merging iteration could represent an emergent topic, potentially displaying considerable divergence from other topics. Thus, merging topics based primarily on their sizes could introduce bias into the process. Therefore, in line with the methodology of agglomerative hierarchical clustering, we advocate for an iterative merging of topics based primarily on their similarity, rather than their respective sizes.

Visualization

Visualizing the trends of HPC adoption in healthcare based on application domain

To visualize trends of HPC adoption in healthcare, we employ four types of graphical representations: Word clouds, stacked area charts, normalized stacked area charts, and violin plots. Before proceeding with visualization, we again utilize the `text-davinci-003` model via the OpenAI API to summarize the topic in less than 10 words based on the top 20 keywords extracted from each topic.

As a variant of the classic area chart, stacked area chart partitions the area into segments, each representing a specific topic. The `x-axis` denotes the publication years, and the `y-axis` represents the cumulative count of publications. The thickness of each segment within a given year directly corresponds to the number of publications for that particular topic. This not only allows for an easy understanding of the distribution of individual topics over time but also visualizes the total volume of publications for each year. Additionally, the normalized stacked area chart effectively highlights the comparative size of each topic within the overall research landscape, enabling the identification of shifts in academic concentration across time. It provides a comprehensive, aggregate perspective of topic prevalence over the years, underlining research trend evolution.

The violin plot serves as another efficient visualization tool for presenting the trends of publication based on topics across different publication years. It merges the attributes of a box plot and a kernel density plot to provide a comprehensive view of the data

[5] https://github.com/ddangelov/Top2Vec.

Table 1 Embedding model evaluation results

Embedding model	Model types	CV coherence
Doc2vec	Doc2vec	**0.622**
Universal sentence encoder large	Universal sentence encoder	0.423
Universal sentence encoder multilingual large	Universal sentence encoder	0.391
All-MiniLM L6-v2	BERT	0.351
Paraphrase-multilingual MiniLM-L12-v2	BERT	0.332

distribution. The 'violins' broadness represents the density of publications, facilitating an intuitive comprehension of the prevalent topics for a specific year. It provides an intuitive comparison of publication activity across multiple topics and years, providing insights into the transformation of scholarly focuses over time.

Visualizing the correlation and convergence of topics

In our study, we capitalize on the UMAP generated by Top2Vec and adapt it to produce an interactive bubble chart to illustrate the correlation and convergence of topics. Given that UMAP reduces the high-dimensional document vectors to a two-dimensional space, the similarity of each document within a given topic is depicted by their proximity in this visual representation. Furthermore, we choose to symbolize the centroid of each topic with a triangle marker and prominently feature the top three most-cited papers from each topic for subsequent reading and analysis. To effectively illustrate the correlation and convergence among various topics, we construct a circle centered at the centroid of each topic, encompassing 70% of the literature associated with the topic. The areas of overlap among these circles serve as indicators of potential convergence across diverse topics.

Results

In the results section, we first compare different embedding models utilized within the top2vec algorithm, highlighting the performance differences and selecting the most fitting model for our data. Secondly, the process of topic merging is delineated through dendrograms, guiding the optimal selection of the number of topics. This is further complemented by visualizing the trends of HPC adoption in healthcare applications through an array of graphical representations including word clouds, stacked area charts, normalized stacked area charts, and violin charts. The correlations and convergence of topics are further explored through bubble charts. providing insights into the inter-relationships between different subject areas. Finally, through all the analysis results mentioned previously, we engage in an in-depth discussion of what we have learned from the past and present of HPC adoption in healthcare to identify its future strategic opportunities. This sets the stage for a comprehensive understanding of the evolution and potential of HPC in the healthcare domain.

Embedding model comparison results

As described in section , we use CV coherence score to evaluate five embedding models. As shown in Table 1, the Doc2Vec model achieves the highest coherence score of 0.622.

Fig. 3 Dendrogram of hierarchical clustering with Identified optimal topic number

This is followed by the Universal Sentence Encoder Large model with a score of 0.423, the Universal Sentence Encoder Multilingual Large model with 0.391, the all-MiniLM-L6-v2 model scoring 0.351, and lastly the paraphrase-multilingual-MiniLM-L12-v2 model with 0.332. Based on these outcomes, the Doc2Vec model is chosen as our default embedding model for Top2Vec due to its superior performance.

Visualizing the process of topic merging using dendrograms for optimal topic number selection

As described in section , we employ a dendrogram to intuitively determine the optimal number of topics by visualizing the merging process of agglomerative hierarchical clustering. As presented in Fig. 3, Top2Vec initially identified 24 topics in the topic modeling phase. The merging points indicate the remaining number of topics after consolidation. Among these, we observe several similar topics with high granularity. Specifically, six topics are related to genomics analysis, two pertain to epidemic and virus research, and two are associated with drug discovery (as shown in the red boxes with dashed lines). To effectively illustrate the primary trends of HPC adoption in healthcare, we merge topics with similar concepts by following the aggregation pathway in the dendrogram. After completing the merging process, we find that the residual number of topics is reduced to 11. Therefore, we identify an appropriate cutoff at these eleven topics (marked by a vertical grey dashed line) and merge the topics following the methodology outlined in section . New keywords for each topic are generated by identifying the n-closest word vectors to the resulting topic vector.

Trends of HPC adoption in healthcare applications

Figure 4 illustrates word clouds for eleven distinct topics, derived from the merging process. Each sub-word cloud's title serves as a summary of the topic words, which are generated by the `text-davinci-003` model as discussed in "Visualizing the trends of HPC adoption in healthcarebased on application domain" section. The identified topics include 'Healthcare ICT infrastructure', 'Genomic Analysis', 'Numerical simulation of blood flow and cardiology', 'Drug Discovery', 'Deep learning medical

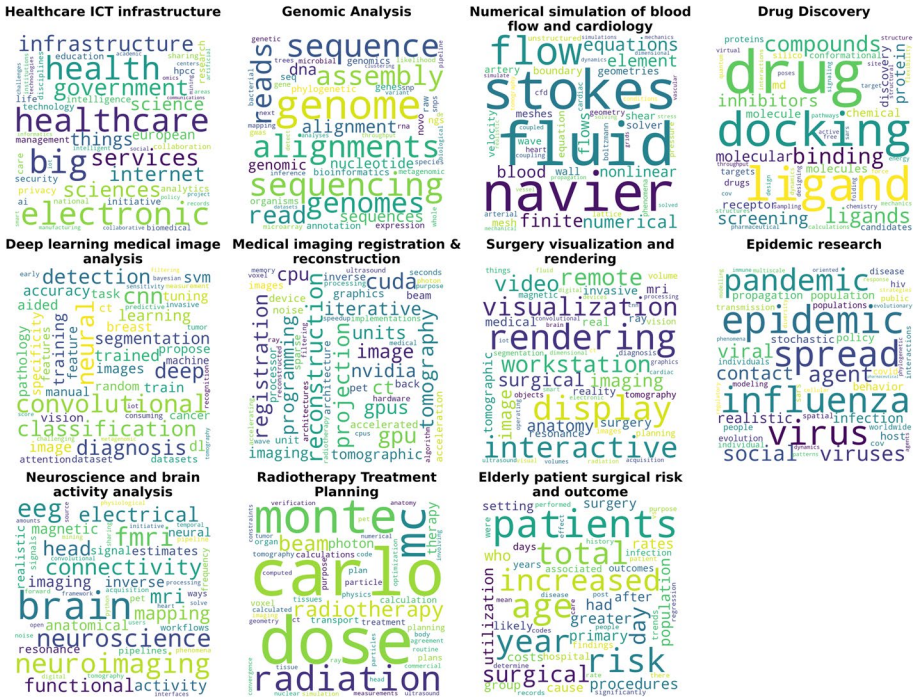

Fig. 4 Word clouds of eleven topics derived from Top2Vec. The main application domains include genomic analysis, drug discovery, medical image analysis, surgery visualization and epidemic research

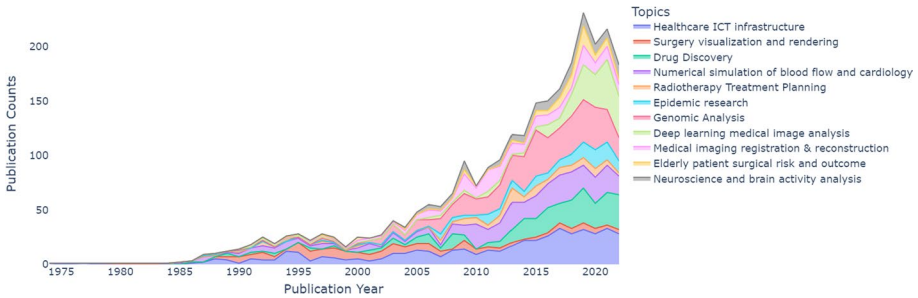

Fig. 5 Stacked area chart illustrating trends in HPC adoption in healthcare cross application domains over time. Each segment represents a specific topic. The thickness of segments reflects topic-specific publication volume

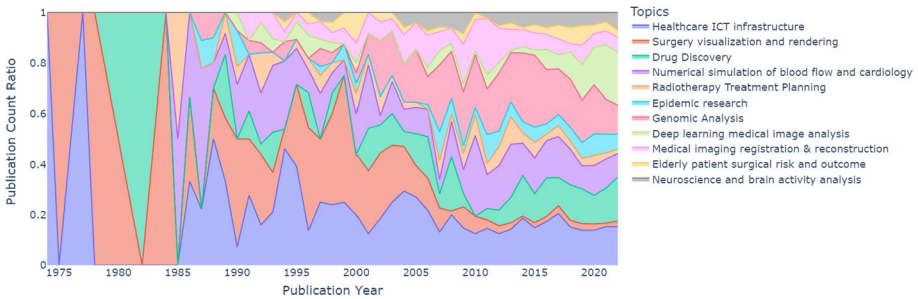

Fig. 6 Normalized stacked area chart depicting the relative distribution of HPC adoption in healthcare across application domains over time. Each segment represents a specific topic. The thickness of segments reflects normalized proportion of publication counts

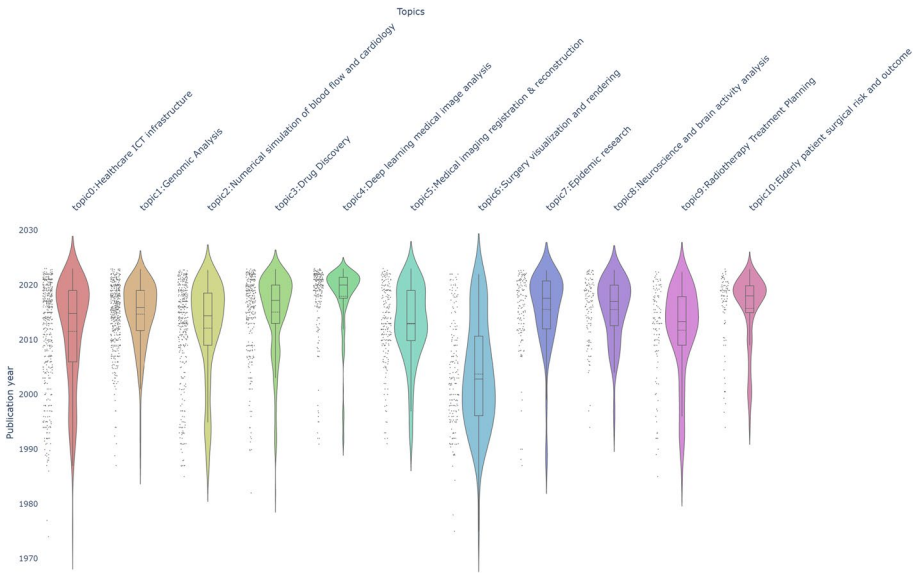

Fig. 7 Violin chart depicting the distribution of HPC adoption in healthcare across application domains over time. Violin width represents publication density

image analysis', 'Medical imaging registration & reconstruction', 'Surgery visualization and rendering', 'Epidemic research', 'Neuroscience and brain activity analysis', 'Radiation Therapy Planning', and 'Elderly patient surgical risk and outcome'.

As indicated by the stacked area chart and normalized stacked area chart in Figs. 5 and 6, the utilization of HPC in healthcare has seen significant growth over the past four decades. Based on the data extracted from Scopus, the pioneering paper highlighting HPC's contribution to healthcare, specifically the rapid analysis of long electrocardiogram (ECG) records, was published in 1974 [58]. Notably, the number of annual publications on this topic triples after year 2020, compared to that of the year 2010. Prior to 2005, the majority of the literature focuses on healthcare ICT infrastructure. However, there has been a marked upsurge in studies pertaining to genomics analysis since 2000. At the end of 1999, IBM announced the Blue Gene project, which was focused on investigating biomolecular phenomena, such as genomics and protein folding, through the use of a petaFLOPS supercomputer [59]. This initiative was recognized as a significant milestone in leveraging supercomputing power to support bioinformatics research. Starting in 2010, the utilization of HPC becomes prevalent in drug discovery efforts. Mak et al. explore the role of AI in drug discovery, notably enhanced by supercomputing, illustrating its widespread use throughout the pharmaceutical product lifecycle [60]. This includes support in drug screening, predicting bioactivity and toxicity, and categorizing drug molecules. The study shows that the application of AI spans the entire spectrum of the pharmaceutical industry, from drug discovery to product management, highlighting its critical role in enhancing pharmaceutical research and development processes. Since 2015, researchers extensively employ HPC in medical image analysis, leveraging deep learning techniques. Additionally, the outbreak of COVID-19 in 2019 triggered a noticeable increase in HPC's application in pandemic-related research.

The violin chart presented in Fig. 7 confirms the insights derived from the stacked area chart, specifically the extensive investigation of healthcare ICT infrastructure in the past thirty years. From 2005 onwards, a marked shift has been observed in the utilization of HPC in medical imaging research. Initially, the focus was predominantly on image registration and reconstruction. However, after 2015, there was a significant transition towards the incorporation of deep learning techniques for medical image analysis. An analysis of the integration of deep learning in medical image examination shows a significant increase in related publications during 2015 and 2016, due to the development of deep convolutional networks [61]. This trend was significantly influenced by Krizhevsky et al.'s contribution to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 with their proposed convolutional neural network (CNN) architecture, AlexNet, which set new standards by winning with an unprecedented margin [62]. Following this, the development of even more sophisticated, deeper network architectures has continued to advance the field, making deep convolutional networks the preferred method in computer vision. Fluid dynamics simulations of blood flow and cardiology, which adopted HPC relatively early, witnessed a significant surge after 2008. Additionally, visualization and rendering in surgical practice, among the earliest applications of HPC, were hot research topics from 1990 to 2010, although interest has waned since. Moreover, research related to epidemic spreading emerged around 2003, concurrent with the SARS-CoV-1 outbreak in Asia, and experienced a tremendous increase in HPC utilization during the global spread of COVID-19, underscoring HPC's growing importance in managing global health crises. The 2021 survey paper highlights a multifaceted role of Deep Learning in helping combat the COVID-19 pandemic through advances in various domains, such as Natural Language Processing, Computer Vision, Life Sciences, and Epidemiology [63]. It details how the application of these technologies differs based on the availability of large datasets and the structuring of learning tasks, and offers guidance for future research directions on COVID-19, emphasizing the need for integrated approaches across disciplines. Another noteworthy observation is the rising popularity of research on risk and treatment outcomes for elderly patients since 2019, possibly correlated to pandemic research, given the heightened concern for elderly individuals during this crisis. Additionally, the exploding healthcare expenses and the increasing average age of the global population further underscore this emphasis.

Visualizations of topics correlation and convergence

A bubble chart is constructed using UMAP to visually represent the coverage area of each topic (Fig. 8). Translucent circles were utilized to depict the extent of coverage, while centroid topic vectors are indicated by triangular markers. Additionally, square markers are employed to symbolize the three most-cited papers for each topic. The top six keywords associated with each topic are also presented beneath their respective vectors. All publications form a point cloud, depicted in the background. The chart also incorporates interactive visualization mechanisms: hovering the mouse cursor over each square marker brings up details such as the paper's associated topics, ID, title, citation count, and abstract. This bubble chart provides substantial insights into the correlation and potential convergence trends of the topics, observable through the distances and overlaps between them. As indicated in Fig. 8, a notable observation is

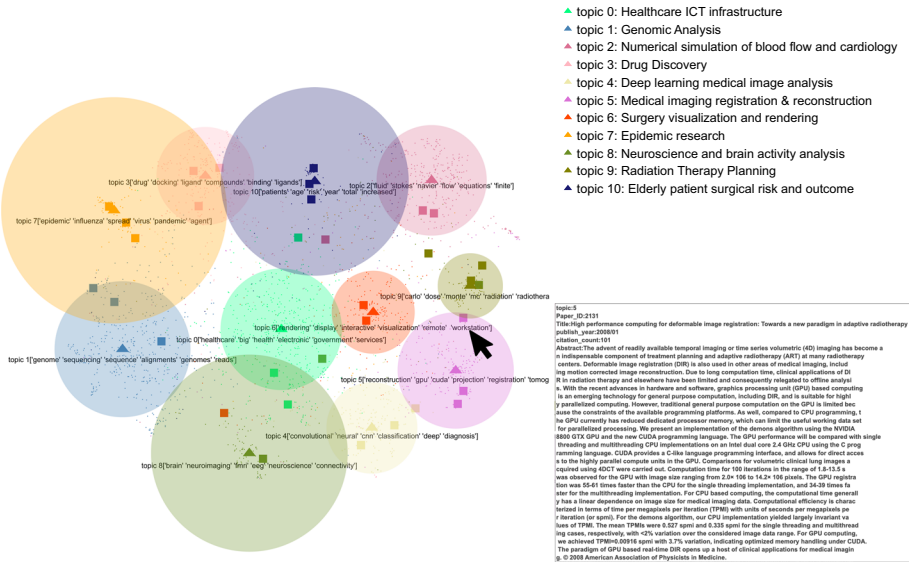

Fig. 8 Bubble chart showing topic coverage areas based on UMAP. Translucent circles represent each topic's coverage area, centered at the topic's centroid and encompassing 70% of associated literature. Centroid topic vectors are represented by triangular markers, and the three most-cited papers are indicated with square markers. A text box provides in-depth details on highly cited papers, including associated topics, ID, title, citation count, and abstract

the proximity of topics related to deep learning for medical image analysis to neuroscience and brain research. This suggests an extensive application of HPC-enabled medical imaging analysis within neuroscience research. The review underscores the growing trend towards utilizing GPU-enabled deep learning for brain MRI segmentation, attributed to its self-learning capabilities and adaptability to large datasets [64]. With ongoing improvements in deep learning frameworks, these methods are increasingly surpassing traditional machine learning techniques, establishing a new benchmark in the domain of medical imaging analysis. In addition, the research related to simulations of cardiovascular & blood flow, epidemic research, and drug discovery closely aligns with and overlap the topic of surgical risk and outcomes for elderly patients, indicating a focus on elderly patient outcomes across these topics. Furthermore, by examining the position of highly-cited papers within the bubble chart, we notice that some papers either represent pioneering studies, or span multiple domains, applying specific techniques to innovative fields. For instance, one of the highly cited papers titled 'High performance computing for deformable image registration: Towards a new paradigm in adaptive radiotherapy' [65] (marked with a black arrow), is located in the intersection area of the topics 'medical imaging registration & reconstruction' and 'radiation therapy planning'. This paper highlights the implementation of HPC for near real-time deformable image registration in radiotherapy, indicating that interdisciplinary papers might contribute to greater popularity and citation rates.

Identifying opportunities through analysis of past and present HPC adoption in healthcare

Historical analysis of HPC adoption within the healthcare sector illustrates a compelling transformation from a specialized technology into a foundational tool for modern medicine. Originally, HPC was predominantly utilized for complex modeling and

biological computations within academic research. However, its role has become more encompassing, opening new and fascinating medical and business opportunities over time. The convergence between simulation and AI, such as neural rendering, seamlessly melds computer graphics and deep learning, opening new avenues in diagnostics and personalized care. Concurrently, the consistent growth in genomics and drug discovery, accelerated by HPC, allows businesses to create dynamic models for pharmaceutical development, reducing cost and time-to-market. In addition, the rise of HPC in global epidemic research presents unique opportunities for developing real-time health monitoring systems, aligning commercial interests with public health needs, and enhancing our response to global health crises. The integration of HPC with emerging technologies such as AI and the Internet of Things (IoT) has further fostered innovative applications in telemedicine and remote monitoring, creating opportunities for healthcare providers to offer services across geographical boundaries. The fusion of technological advancements with unique patient needs has the potential to give birth to virtual care platforms, personalized treatment protocols, or even community-based wellness programs. These innovations align with the growing focus on aging populations, offering a compassionate and tailored approach to healthcare that could become a promising business domain.

The landscape of HPC adoption in healthcare represents an invitation to visionary business thinking, reflecting a future where innovation meets well-being. Collaboration among technology vendors, healthcare institutions, pharmaceutical companies, and startups can leverage the lessons learned from the past and the current state of HPC technology to forge partnerships, develop new products, and create service models that cater to an increasingly interconnected and data-driven healthcare environment. The analytical understanding of this evolution can serve as a guide for future endeavors, nurturing an ecosystem where technological advancement is aligned with healthcare requirements. In this convergence, both sectors stand to gain substantially from sustained cooperation and exploration. More importantly, as regulatory landscapes evolve to embrace technological advancement, it is imperative for stakeholders to understand the historical growth and current trends in HPC adoption in healthcare, as it will guide strategic decision-making and foster further innovation and growth in this interdisciplinary field.

Discussion

This study presents an automatic literature analysis framework that utilizes advanced topic modeling techniques (Top2Vec), within the context of the field of HPC adoption in healthcare as the testbed. Through this framework, we effectively process and analyze a substantial volume of scientific literature. The interactive visualizations shed light on the key trends and recognized the emerging research directions in a highly efficient manner. Given that the landscape of HPC adoption in healthcare is extensive and rapidly evolving, the automatic literature analysis offers a scalable and practical approach for researchers and stakeholders to identify trends and potential avenues for future exploration. Continuous tracking of these developments is necessary to maintain a dynamic understanding of the HPC landscape in healthcare, which enables the anticipation of its future.

One of the main limitations of using topic models, such as Top2Vec is the necessity for human intervention to determine the optimal number of topics, which is the only step requiring human input in our otherwise fully automatic literature analysis pipeline. Although metrics such as perplexity and coherence scores can provide guidance towards an optimal topic number, existing literature suggests that these metrics may occasionally be misleading and do not always align with human judgement [66, 67]. Determining the ideal number of topics within a topic model is a subjective decision that depends on the specific context and goals of the analysis. It requires a balance between having a sufficient number of topics to capture the underlying themes in the data and avoiding excessive fragmentation or overlap between topics. Therefore, human experience and judgment play a critical role in defining the optimal topic count. For this reason, we have integrated the use of a dendrogram to assist in determining the optimal topic count in a more intuitive manner.

Another challenge in using topic modeling is related to semantic understanding. While topic models are highly effective in identifying word and document patterns, they fundamentally lack the ability to understand the meaning of the words. This limitation can lead to topics that are difficult to interpret or exhibit semantic inconsistencies, making manual review and interpretation necessary. One possible approach to tackle this challenge is through EL, which can enhance topic modeling by providing an additional layer of explainability [41, 68]. By linking entities to a knowledge base, EL provides contextual understanding, which refines the interpretability of the topics. It effectively resolves ambiguities where identical names might refer to distinct entities, thereby improving the precision of topic clusters. Moreover, while topic modeling can significantly aid in analyzing the vast volumes of scientific literature, it cannot replace the critical and contextual understanding that researchers bring to literature review. It should be regarded as a supplementary tool designed to assist in guiding literature exploration rather than serving as an autonomous solution. Given these limitations, future research should strive to fine-tune this methodology, improve the interpretability of the resulting topics, and investigate alternative methods for model evaluation. Despite the inherent challenges, topic modeling remains a potent instrument for managing and comprehending the ever-expanding corpus of scientific literature.

In conclusion, we propose an automatic literature analysis framework which can be easily applied across diverse literature domains, exemplifying its utility through the examination of literature within the field of HPC adoption in healthcare. The insights derived from this study are expected to guide researchers and practitioners toward recognizing emerging opportunities and challenges in the deployment of HPC in healthcare. These findings would contribute to a more informed and strategic incorporation of HPC in healthcare settings, holding the potential to transform medical research and clinical practice in the years to come.

Abbreviations

AI	Artificial intelligence
CNN	Convolutional neural network
ECG	Electrocardiogram
EHRs	Electronic health records
EL	Entity linking
HDBSCAN	Hierarchical density-based spatial clustering of applications with noise

HPC	High performance computing
IoT	Internet of things
LDA	Latent Dirichlet Allocation
LSA	Latent semantic analysis
NMF	Non-negative matrix factorization
NPMI	Normalized pointwise mutual information
UMAP	Uniform manifold approximation and projection

Acknowledgements
Not applicable.

Author contributions
JL and SW: Methodology, data preprocessing, visualization, validation, manuscript drafting. SR and AO: Conceptualization, methodology, supervision, reviewing & editing. All authors reviewed the manuscript and approved the final draft.

Funding
Financial support for this study was provided in part by Atos through the HPC, AI and Quantum Life Sciences Centre of Excellence (CEPP); as well as by SURF, the collaborative organization for IT in Dutch education and research. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Jieyi Li, Shuai Wang, Stevan Rudinac, Anwar Osseyran, declare that they do not have any financial or personal relationships with other people or organizations that could have inappropriately influenced this study.

Received: 14 September 2023 Accepted: 22 April 2024
Published online: 02 May 2024

References

1. Elsebakhi E, Lee F, Schendel E, Haque A, Kathireason N, Pathare T, Syed N, Al-Ali R. Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. J Comput Sci. 2015;11:69–81.
2. Raj P, Raman A, Nagaraj D, Duggirala S, Raj P, Raman A, Nagaraj D, Duggirala S. Big data analytics for healthcare. High-performance big-data analytics: computing systems and approaches. 2015;391–424.
3. Jia X, Ziegenhein P, Jiang SB. Gpu-based high-performance computing for radiation therapy. Phys Med Biol. 2014;59(4):151.
4. Bastrakov S, Meyerov I, Gergel V, Gonoskov A, Gorshkov A, Efimenko E, Ivanchenko M, Kirillin M, Malova A, Osipov G, et al. High performance computing in biomedical applications. Procedia Comp Sci. 2013;18:10–9.
5. Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. Drug Discov Today. 2017;22(4):712–7.
6. Stöcker T, Vahedipour K, Pflugfelder D, Shah NJ. High-performance computing MRI simulations. Magn Reson Med. 2010;64(1):186–93.
7. Alanazi HO, Zaidan A, Zaidan B, Kiah MM, Al-Bakri S. Meeting the security requirements of electronic medical records in the era of high-speed computing. J Med Syst. 2015;39:1–13.
8. Vitabile S, Marks M, Stojanovic D, Pllana S, Molina JM, Krzyszton M, Sikora A, Jarynowski A, Hosseinpour F, Jakobik A, et al. Medical data processing and analysis for remote health and activities monitoring. 2019;186–220.
9. Molidor R, Sturn A, Maurer M, Trajanoski Z. New trends in bioinformatics: from genome sequence to personalized medicine. Exp Gerontol. 2003;38(10):1031–6.
10. Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, Blayney JK. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform. 2019;20(5):1795–811.
11. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241–50. https://doi.org/10.1016/j.drudis.2018.01.039.
12. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.

13. Zhang X, Wong SE, Lightstone FC. Toward fully automated high performance computing drug discovery: a massively parallel virtual screening pipeline for docking and molecular mechanics/generalized Born surface area rescoring to improve enrichment. ACS Publications. 2014.
14. Ge H, Wang Y, Li C, Chen N, Xie Y, Xu M, He Y, Gu X, Wu R, Gu Q, et al. Molecular dynamics-based virtual screening: accelerating the drug discovery process by high-performance computing. J Chem Inf Model. 2013;53(10):2757–64.
15. Sanbonmatsu K, Tung C-S. High performance computing in biology: multimillion atom simulations of nanoscale systems. J Struct Biol. 2007;157(3):470–80.
16. Kharche S, Seemann G, Margetts L, Leng J, Holden AV, Zhang H. Simulation of clinical electrophysiology in 3d human atria: a high-performance computing and high-performance visualization application. Concurr Comput Pract Exp. 2008;20(11):1317–28.
17. Perrin D, Ruskin HJ, Crane M. Model refinement through high-performance computing: an agent-based hiv example. In: Immunome Research, vol. 6, pp. 1–9. BioMed Central; 2010.
18. Phong TD, Duong HN, Nguyen HT, Trong NT, Nguyen VH, Van Hoa T, Snasel V. Brain hemorrhage diagnosis by using deep learning. In: Proceedings of the 2017 International Conference on Machine Learning and Soft Computing. 2017;pp. 34–39.
19. Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol. 2019;58:161–7.
20. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
21. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. Ann Trans Med. 2020;8(11).
22. Tahmassebi A, Gandomi AH, McCann I, Schulte MH, Goudriaan AE, Meyer-Baese A. Deep learning in medical imaging: fmri big data analysis via convolutional neural networks. In: Proceedings of the Practice and Experience on Advanced Research Computing. 2018; pp. 1–4.
23. Lee H, Turilli M, Jha S, Bhowmik D, Ma H, Ramanathan A. Deepdrivemd: Deep-learning driven adaptive molecular simulations for protein folding. In: 2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS), IEEE; pp. 12–19. 2019.
24. Bai Q, Liu S, Tian Y, Xu T, Banegas-Luna AJ, Pérez-Sánchez H, Huang J, Liu H, Yao X. Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(3):1581.
25. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84.
26. Jacobi C, Van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. 2018;89–106.
27. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
28. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
29. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. JASIST. 1990;41(6):391–407.
30. Alghamdi R, Alfalqi K. A survey of topic modeling in text mining. Int J Adv Comput Sci Appl(IJACSA). 2015;6(1).
31. Yi X, Allan J. A comparative study of utilizing topic models for information retrieval. In: Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31, Springer; pp. 29–41. 2009.
32. Meeks E, Weingart SB. The digital humanities contribution to topic modeling. JDH. 2012;2(1):1–6.
33. Asmussen CB, Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 2019;6(1):1–18.
34. Amado A, Cortez P, Rita P, Moro S. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. Eur Res Manag Bus Econ. 2018;24(1):1–7.
35. Chen H, Wang X, Pan S, Xiong F. Identify topic relations in scientific literature using topic modeling. IEEE Trans Eng Manag. 2019;68(5):1232–44.
36. Älgå A, Eriksson O, Nordberg M. Analysis of scientific publications during the early phase of the covid-19 pandemic: topic modeling study. J Med Internet Res. 2020;22(11):21559.
37. Lindstedt NC. Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. Soc Curr. 2019;6(4):307–18.
38. Altena AJ, Moerland PD, Zwinderman AH, Olabarriaga SD. Understanding big data themes from scientific biomedical literature through topic modeling. J Big Data. 2016;3(1):1–21.
39. Maier D, Waldherr A, Miltner P, Wiedemann G, Niekler A, Keinert A, Pfetsch B, Heyer G, Reber U, Häussler T, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. Commun Methods Meas. 2018;12(2–3):93–118.
40. Angelov D. Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470. 2020.
41. Rudinac S, Gornishka I, Worring M. Multimodal classification of violent online political extremism content with graph convolutional networks. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. Thematic Workshops '17, pp. 245–252. Association for Computing Machinery, New York, NY, USA; 2017. https://doi.org/10.1145/3126686.3126776.
42. Egger R, Yu J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Front Sociol. 2022;7.
43. Karas B, Qu S, Xu Y, Zhu Q. Experiments with lda and top2vec for embedded topic discovery on social media data-a case study of cystic fibrosis. Front Artif Intell. 2022;5.
44. Le Q, Mikolov T. Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196. PMLR; 2014.

45. Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, et al. Universal sentence encoder for english. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018; pp. 169–174.
46. Yang Y, Cer D, Ahmad A, Guo M, Law J, Constant N, Abrego GH, Yuan S, Tar C, Sung Y-H, et al. Multilingual universal sentence encoder for semantic retrieval. 2019. arXiv preprint arXiv:1907.04307.
47. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. 2019. arXiv preprint arXiv:1908.10084.
48. Reimers N, Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation. 2020. arXiv preprint arXiv:2004.09813.
49. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015; pp. 399–408.
50. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arxiv 2018. arXiv preprint arXiv:1802.03426. 1802.
51. Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 160–172. Springer; 2013.
52. Zografos G, Moussiades L. A gpt-based vocabulary tutor. In: International Conference on Intelligent Tutoring Systems, pp. 270–280. Springer; 2023.
53. Carpenter KA, Altman RB. Using gpt-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules. 2023;13(2):387.
54. Bommarito II M, Katz DM. Gpt takes the bar exam. 2022. arXiv preprint arXiv:2212.14402.
55. Nielsen F, Nielsen F. Hierarchical clustering. Introduction to HPC with MPI for Data Science. 2016;195–211.
56. Orkphol K, Yang W. Word sense disambiguation using cosine similarity collaborates with word2vec and wordnet. Future Internet. 2019;11(5):114.
57. Rozado D. Using word embeddings to analyze how universities conceptualize "diversity" in their online institutional presence. Society. 2019;56:256–66.
58. Clark KW, Nolle FM, Cox JR, Oliver GC. High performance computer programs for rapid analysis of long ecg records. In: San Diego Biomed Symp, Proc; 1974.
59. Allen F, Almasi G, Andreoni W, Beece D, Berne BJ, Bright A, Brunheroto J, Cascaval C, Castanos J, Coteus P, et al. Blue gene: a vision for protein science using a petaflop supercomputer. IBM Syst J. 2001;40(2):310–27.
60. Mak K-K, Wong Y-H, Pichika MR. Artificial intelligence in drug discovery and development. Drug Discov Eval 2023;1–38.
61. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
62. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25.
63. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. J Big Data. 2021;8(1):1–54.
64. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging. 2017;30:449–59.
65. Samant SS, Xia J, Muyan-Özçelik P, Owens JD. High performance computing for deformable image registration: towards a new paradigm in adaptive radiotherapy. Med Phys. 2008;35(8):3546–53.
66. Hasan M, Rahman A, Karim MR, Khan MSI, Islam MJ. Normalized approach to find optimal number of topics in latent dirichlet allocation (lda). In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020, Springer; pp. 341–354. 2021.
67. Harrando I, Lisena P, Troncy R. Apples to apples: A systematic evaluation of topic models. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021; pp. 483–493.
68. Dillan T, Fudholi DH. Ldaviewer: An automatic language-agnostic system for discovering state-of-the-art topics in research using topic modeling, bidirectional encoder representations from transformers, and entity linking. IEEE Access; 2023.

Publisher's Note