scientific reports

Check for updates

OPEN

Evolution of AI enabled healthcare systems using textual data with a pretrained BERT deep learning model

Yi Jie Wang[1], Wei Chong Choo[1,2], Keng Yap Ng[2,3], Ran Bi[4] & Peng Wei Wang[5]✉

In the rapidly evolving field of healthcare, Artificial Intelligence (AI) is increasingly driving the promotion of the transformation of traditional healthcare and improving medical diagnostic decisions. The overall goal is to uncover emerging trends and potential future paths of AI in healthcare by applying text mining to collect scientific papers and patent information. This study, using advanced text mining and multiple deep learning algorithms, utilized the Web of Science for scientific papers (1587) and the Derwent innovations index for patents (1314) from 2018 to 2022 to study future trends of emerging AI in healthcare. A novel self-supervised text mining approach, leveraging bidirectional encoder representations from transformers (BERT), is introduced to explore AI trends in healthcare. The findings point out the market trends of the Internet of Things, data security and image processing. This study not only reveals current research hotspots and technological trends in AI for healthcare but also proposes an advanced research method. Moreover, by analysing patent data, this study provides an empirical basis for exploring the commercialisation of AI technology, indicating the potential transformation directions for future healthcare services. Early technology trend analysis relied heavily on expert judgment. This study is the first to introduce a deep learning self-supervised model to the field of AI in healthcare, effectively improving the accuracy and efficiency of the analysis. These findings provide valuable guidance for researchers, policymakers and industry professionals, enabling more informed decisions.

Keywords Healthcare, Text mining, BERT, Technology management, AI

With the rapid development of digital technology and the internet, information technology with Artificial Intelligence (AI) has become a key force driving reform, innovation and sustainable development across various industries in developing economies. Since John McCarthy first proposed the concept of AI in 1956, the field has undergone significant evolution[1]. Initially limited to pattern recognition applications, AI has developed into multiple subfields, such as machine learning, computer vision, and natural language processing (NLP), continuously broadening its application scope. Through the integration and optimisation of these technologies, traditional industries can be revived through technological innovation and development, thereby promoting sustainable economic growth and improving social welfare[2]. Taking healthcare as an example, AI has been used to assist medical decision-making as early as the 1980s to improve medical efficiency and social benefits[3]. With the advancement of technology, especially the development of computing power and big data, AI has been widely applied in many fields such as document management, medical imaging, genomics and drug research development, greatly improving the quality of diagnosis and treatment. These illustrate the significant potential of AI in healthcare, especially in improving diagnostic accuracy, optimising treatment plans, and enhancing patient management. However, the rapid iteration and updating of current technology presents a series of challenges, and the stages and trends of the technology life cycle have not yet been clearly defined[4]. Meanwhile, the specific trends and impacts of how AI will shape the healthcare market remain unclear[2]. The rapid technological iteration creates decision-making complexity, especially as the challenges facing the

[1]School of Business and Economics, Universiti Putra Malaysia, Seri Kembangan, Malaysia. [2]Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, Seri Kembangan, Malaysia. [3]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Seri Kembangan, Malaysia. [4]SAS Institute Inc., 100 SAS Campus Drive, Cary, USA. [5]School of Physics and Electronic Information, Jiangsu Second Normal University, Nan Jing, Jiang Su, China. ✉email: pengweiwang2024@jssnu.edu.cn

public health system become increasingly apparent. These challenges, including uneven resource distribution, inadequate emergency response mechanisms, and insufficient public health infrastructure urgently require in-depth research and effective solutions utilizing advanced technologies such as AI. Furthermore, addressing the sustainable development issues and filling gaps in current healthcare market research require skilfully identifying technology opportunities in the market and investigating the development and innovation trends of AI in healthcare through technology life cycle analysis.

This research aims to identify the core themes, technology evolution and commercialisation process of AI in the healthcare industry; analyse and explore the strategic innovation and market dynamics of AI in healthcare; and reveal its development trends and market prospects. The research explores the following questions: What are the development trends of AI technology innovation in the healthcare field? Which subfields of AI will become short-term market investment hot spots?

This study employs text mining techniques to analyse key scientific papers and patents on AI in healthcare, particularly by utilising BERT to explore emerging trends and their potential market opportunities. In addition to filling the gaps in existing research, it also attempts to provide decision-making support for policymakers and industry practitioners to promote healthy development and application of AI in healthcare.

Literature review

In recent years, research on information technology trends and emerging technologies has attracted widespread attention in the healthcare industry, with a significant increase in the number of documents, patents and related investments. Electric healthcare, mobile healthcare and telemedicine in the field of healthcare are closely related to digital technologies, such as AI, Internet of Things (IoT), information and communication technology (ICT) and many more[5]. These technologies are used to improve industry efficiency, productivity and quality to provide higher quality of care. AI, as the most representative emerging information technology, uses algorithms to perform intelligent tasks[6]. To some extent, AI can significantly reduce human workload and improve the efficiency of processing complex tasks. According to the systematic review in Fig. 1, the main theoretical models used in the relevant AI literature are as follows[7]. Many scholars apply methods and theories from the health field to study the social process of technology, including adoption, diffusion and institutionalisation. For example, the technology acceptance model (TAM), theory of planned behaviour (TPB), and Unified Theory of Acceptance and Use of Technology (UTAUT) provide a solid theoretical foundation for understanding and predicting the adoption of AI in healthcare[8–10]. Firstly, these theoretical models emphasise the influence of expected utility, ease of use, social influence, and convenience on technology adoption intentions. Secondly, methods such as the Super-SBM-logit model (applied at the economic level) and the technology–organisation–environment framework (TOE) (applied at the organisational level) have also been supported by empirical research in healthcare[7]. These studies not only analyse the economic benefits of technology adoption, but also consider the impact of organisational culture and the external environment on technology adoption. However, most current research focuses on the adoption of AI technology and specific AI applications, such as diagnostic assistance systems, with less emphasis on the development trends and potential market opportunities of AI in healthcare. The possible future development directions and technology life cycle are still unclear.

Effectively capturing potential technology opportunities in the market and using the technology life cycle to explore the emerging AI technology trends in healthcare are key to addressing the main points and gaps in current healthcare market research. Some scholars have reviewed the scientific literature on information technology to understand the evolution of digital technology development in healthcare research[11–13]. To be more specific, with the sudden increase in data samples in the healthcare industry and the emergence of text information such as medical record reports, medical papers, and patient feedback, traditional analysis methods to handle this type of data. With the development of text mining technology, some studies have gradually begun to use this method struggle to analyse text data and technology development trends, but the number of studies remains limited[14]. Text mining usually refers to the process of extracting useful information from large amounts of data using NLP methods[15]. For example, identifying cutting-edge technologies in specific fields through text mining and algorithmic clustering can support the development of technology roadmaps and life cycles[14]. Additionally, some studies have identified topics with positive sentiments by analysing comments about information technology posted by medical professionals in online forums. These topics were then used to identify and clarify the role of information technology in improving image diagnostic accuracy and enhancing the potential of

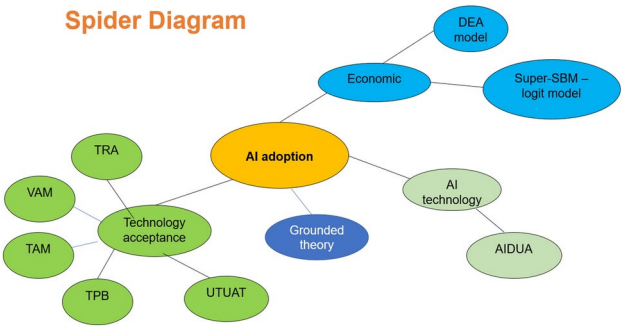

Fig. 1. theories for research on AI adoption[7].

personalised medicine potential[16]. Text mining can reveal patterns, trends and potential correlations within text data in healthcare. In order to better understand technological development trends, technology forecasting has emerged and evolved, reflecting new or rising technological changes through predictions of characteristics, intensity and timing, while seeking to discover and integrate the underlying principles of technology[17]. Some scholars believe that technology forecasting may affect the possibility and performance of business development and have proposed a patent-based topic modelling method based on technology life cycle theory[18]. Therefore, after gaining a better understanding of the emergence of these technologies, predicting their future development trends to meet the needs of technological development, health industry development and market investment is an area requiring further research in the academic community.

At this stage, mainstream academic research focuses on technology research and development and industrial integration, with limited research on technology forecasting in healthcare[19–22]. With the continuous advancement of text mining technology and the improvement in the accuracy of algorithms and deep learning models, these research method innovations have effectively addressed current academic research needs[15,23]. It encompasses a range of techniques and methods, including information retrieval, text classification, text clustering, sentiment analysis, concept extraction and so on[24–26]. Traditional methods for technological identification and forecasting primarily depend on expert annotations, a process not only time-consuming but also vulnerable to personal biases[15,26–28]. Manually extracting unstructured information is time-consuming, resource-intensive and relies on the interpretation of domain experts[29]. As technology becomes more complex and the high volume of structured and unstructured data increases, it becomes increasingly difficult to analyse the value of patents based solely on expert knowledge[18]. For example, structural topic modelling has been used to analyse unsupervised models and conduct forecasting with a focus on the spread and adoption of AI[30]. Although it focuses on adoption, its novelty and innovation in text mining technology and machine learning methods are noteworthy and warrant further study[14,16,26]. Consequently, these text mining methods could be used to analyse text data of technological development and predict future trends, which deserves further study.

In order to address the limitation of traditional analysis methods that are too subjective, deep learning models have gradually become a potential analysis method for technology prediction[27,28,31–34]. BERT, as a newly introduced model in the field of NLP, can reveal complex patterns hidden in text data based on the transformer model, which greatly improves the accuracy of information extraction and text analysis[4,35,36]. Papers, patents, research reports and other forms of text data have become key resources for obtaining information on the latest scientific technological progress and market trends. They can be considered important indicators for identifying technological paths and market trends. The research gap addressed by this study lies in the limited exploration of technology forecasting within the healthcare sector, particularly using advanced text mining techniques and deep learning models. While mainstream academic research has focused on technology research and development, as well as industrial integration, there is a lack of comprehensive studies that apply modern data analysis methods to predict future trends in healthcare technologies. Additionally, existing models often rely on a single source of textual data, such as scientific papers or patents, which limits the scope and depth of insights. This study aims to fill this gap by integrating academic literature and patent information, leveraging advanced models to enhance the accuracy and objectivity of technology forecasting and offering a more holistic approach to forecasting technological developments in healthcare. Hence, this study focuses on and combines two different type of text data: academic literature and patent information, based on pre-trained deep leaning models.

Methods

Text mining

Text mining, which involves extracting useful information and knowledge from large volumes of textual data, employs NLP and machine learning tools and techniques for information retrieval to process and analyse unstructured text[15,26,28]. The principal purpose of text mining is to capture and analyse all possible meanings embedded in the text[15,26,28]. This approach can uncover hidden patterns and trends in a variety of unstructured data sources, including academic literature, patents, news articles, social media posts, Twitter feeds and video transcripts. By employing text mining on scientific papers and patents to identify early indicators of technology trends, promising technologies within specific domains can be identified, supporting technology road mapping and life cycle analysis[14,23]. Traditional topic models such as Latent Dirichlet Allocation (LDA) may generate imprecise or scattered topics due to the subjective judgment of researchers when dealing with highly heterogeneous data sets[18,37].

The bag-of-words model

Under the topic of AI and healthcare, quantitative methods for text data play a vital role in enhancing the explanatory and predictive power of models. By converting text into quantitative indicators, high-quality information can be extracted as new features or variables to explore and solve problems that were previously impossible to quantify and analyse in more depth. In each document, unique words are treated as individual features, and these word occurrences can be used for documents comparison, measuring similarities of documents, topic modelling and text mining[38]. If there are K unique words that are not repeated in M documents, an $M * K$ matrix is formed[38]. And the specific calculation equation is

$$f_{i,j} = \sum_{k}^{m_{i,j}} m_{k,j}$$

The bag-of-words model, while successful in information retrieval tasks, has notable limitations[39]. It fails to capture context or word order, treats words as independent entities ignoring their combined associations, and

can struggle with overly complex models due to an excessive number of features, which may dilute important terms and complicate computations[39].

Bidirectional encoder representations from transformers (BERT)

To address the limitations of traditional, overly subjective analysis methods, deep learning models have gradually become a potential analysis method for technology forecasting[27,28,31–33]. In 2018, BERT was first introduced by Google researchers as a new language representation model[35]. It is the first transformer model to truly take advantage of bidirectional training. This enables the model to better understand the language context. The core innovation of BERT is its use of a large-scale corpus for pre-training, followed by fine-tuning for specific tasks to improve performance across various NLP applications. Unlike traditional topic modelling techniques, such as LDA, which assumes a fixed distribution of topics across the documents, BERT's deep learning-based approach allows for a more nuanced understanding of the text by leveraging its bidirectional attention mechanism. This helps capture intricate dependencies between words in a sentence and across documents. Additionally, compared to t-SNE or PCA, which are primarily used for dimensionality reduction and visualization, BERT excels in extracting meaningful representations from large datasets without requiring explicit topic hypothesis[40,41]. The pre-training process of BERT includes the learning tasks of masked language model (MLM) and next sentence prediction (NSP)[33,36]. During the training process, MLM randomly masks certain words in the input sentence and replaces the original words with special mask labels[29]. Each time a sequence is sampled, MLM randomly masks a certain proportion of words and lets the model predict each masked word-piece label independently[25,29,29,35]. Essentially, a deep learning task is formulated from a large unlabelled corpus of images and the convolutional neural network predicts a masked area of the image for most pretext tasks or predicts the correct angle by which the image is rotated[42]. NSP is a supervised learning task that takes a pair of sentences as input[29]. BERT predicts whether the given two sentences are coherent. This task is achieved by labelling the input sentence pairs (yes or no). The model then learns the relationships between sentences[29]. Supervised learning involves a training process where both the observed data and its corresponding ground truth labels (sometimes referred to as "targets") are essential for training the model[43]. For instance, in breast X-ray examinations, cancerous areas are precisely outlined (as labels), enabling the algorithm to "learn" the characteristics of malignant tumours from these annotated markers[6,31]. In contrast, BERT may learn and complete pre-training through unsupervised learning, where the training data does not have diagnostic or normal/abnormal labels[43].

BERT, as a newly introduced model in the field of NLP, can reveal complex patterns hidden in text data based on the transformer model, which greatly improves the accuracy of information extraction and text analysis[4,35,36]. Papers, patents, research reports and other forms of text data have become key resources for obtaining the latest scientific and technological progress and market trends. It can be considered an important indicator for identifying technological paths and market trends. Due to the complexity and rapid development of technology, analysing only the technology opportunities identified from a single data source may be one-sided. This research aims to use the Colab tool to perform BERT analysis to identify the gap between patents and academic papers, thereby predicting changes in technology development trends.

Data collection and data analysis

To ensure a rigorous and transparent data collection process, this research adheres to PRISMA guidelines, systematically selecting the Web of Science (WOS) for scientific literature and the Derwent Innovations Index (DII) for patent data. WOS was chosen as the source of scientific papers related to AI diagnosis because it includes top interdisciplinary journals and conference papers, and the peer-reviewed, high-quality research content ensures the authority of the research[14]. This is very valuable for tracking the research dynamics of AI in the field of healthcare diagnosis and identifying key research. At the same time, DII is selected as the source of patent data mainly owing to its detailed global patent information, including invention descriptions, legal status and patent citations[44]. In WOS, this research conducted a comprehensive search using keywords such as "artificial intelligence diagnosis" and "healthcare", screening out 1587 papers. In DII, a patent search was conducted through the combination of keywords "artificial intelligence diagnosis", "AI diagnosis", "artificial intelligence", "diagnosis" and "healthcare", yielding 1314 qualifying patents. These data collection activities concluded on June 16, 2023. According to the data collected by WOS and DII document libraries, the unstructured datasets of AI in diagnosis are relatively concentrated from 2018 to 2022. Specifically, there was little relevant text data before 2018, and there was an 18-month intellectual property protection period for the data. Following PRISMA, this study conducted a structured screening process: identification, duplicate removal, full-text review, and final inclusion. To enhance dataset construction, we applied the BERT model to integrate literature and patent data, ensuring a systematic, reproducible, and comprehensive analysis of AI's role in healthcare diagnostics.

After data collection is complete, further data preprocessing is usually performed to improve data quality and validity. This preprocessing process covers key steps such as data cleaning, standardisation, entity recognition and word segmentation. Subsequently, BERT is trained on the data to further analyse AI in healthcare. This begins with model parameter tuning, which involves setting the appropriate learning rate, batch size and number of training epochs to optimise training results. Using the pandas package, the code preprocesses a DataFrame by extracting and formatting dates, cleaning splitting text columns ('TI', 'AB', 'PI') and removing rows with missing or duplicate values. It then filters the data by a specified date range, retains relevant columns, and outputs the cleaned DataFrame. Selecting an appropriate validation set is crucial, as it helps monitor model training progress and prevent overfitting. The next step involves model optimisation. The back propagation algorithm is used to continuously adjust the model weights to ensure that the model can effectively process medical text data. Finally, the model is pre-trained on a large-scale corpus to learn the basic characteristics of language, and then fine-tuned for specific medical tasks to improve task performance and accuracy. In this research, a pre-trained BERT model is used and fine-tuned using the MLM and NSP training tasks. Hyperparameters such as learning

rate, batch size, and the number of training epochs are adjusted during fine-tuning. The code uses UMAP for dimensionality reduction, setting parameters such as n_neighbors, n_components, min_dist and metric. The code implements the K-Means clustering algorithm to group the text embeddings into 50 clusters.

In the current case analysis, this research applies BERTopic to deeply explore the latest trends of AI in the field of healthcare. The flow chart is as follows. This process takes the embedding of the document as the starting point and uses BERT, a bidirectional encoder representation technology, to convert the text into a dense vector form[45]. Using BERT or similar techniques, text data is encoded into vectors containing rich semantic information. This transformation not only captures the complex connections between words but also preserves their meaning in specific contexts. BERTopic uses Sentence Transformers to perform this step, and the specific operation steps will be implemented through Colab's code, as shown in Fig. 2[45].

Document embedding converts text into vectors for machine learning models[45–47]. This study then applies the uniform manifold approximation and projection (UMAP) algorithm to reduce vector dimensionality, while preserving the original data structure[48]. UMAP was chosen over other dimensionality reduction techniques (such as PCA and t-SNE) mainly because UMAP is better at preserving local and global structures when dealing with high-dimensional data, especially for complex relationships in text data[45,48–51]. UMAP efficiently compresses high-dimensional data into lower dimensions by constructing a weighted graph, maintaining local and global structures [45,48–51]. This reduction step enhances computational efficiency and simplifies data visualisation for clustering algorithms.

$$w\left(x_i, x_j\right) = exp\left(\frac{-max\left(0, d\left(x_i, x_j\right) - p_i\right)}{\sigma_i}\right),$$

where $w\left(x_i, x_j\right)$ is the weight of the edge between points x_i and x_j, indicating the similarity between them; $d\left(x_i, x_j\right)$ represents the distance between points x_i and x_j; p_i is the local scale parameter of point x_i , usually representing the distance between the kth nearest neighbour and point x_i, to ensure that each point is connected to at least one other point; σ_i is a length scale parameter used to adjust the influence of distance; $max\left(0, d\left(x_i, x_j\right) - p_i\right)$ is usually used to ensure that only part of the distance beyond the local scale p_i is calculated, thereby limiting the influence of farther points on the weight[50].

After dimensionality reduction, the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm is employed for clustering, identifying natural clusters and handling noisy data without presetting the number of clusters[52]. HDBSCAN calculates the core distance to the kth nearest neighbour and uses mutual reachable distance to differentiate dense and sparse regions, improving the accuracy of clustering[53]. While HDBSCAN identified approximately 10% of the data points as outliers, based on expert judgment, this research determined that in the context of our research dataset, patent data should not be considered outliers. Given this, K-means is a more suitable algorithm for this research. It was used to identify the critical research domains and future research directions[54]. The next step involves creating topic representations using class-based term frequency–inverse document frequency (C-TF–IDF). It identifies keywords for each cluster by calculating their term frequency and inverse document frequency, enhancing topic clarity[46]. Maximum marginal relevance

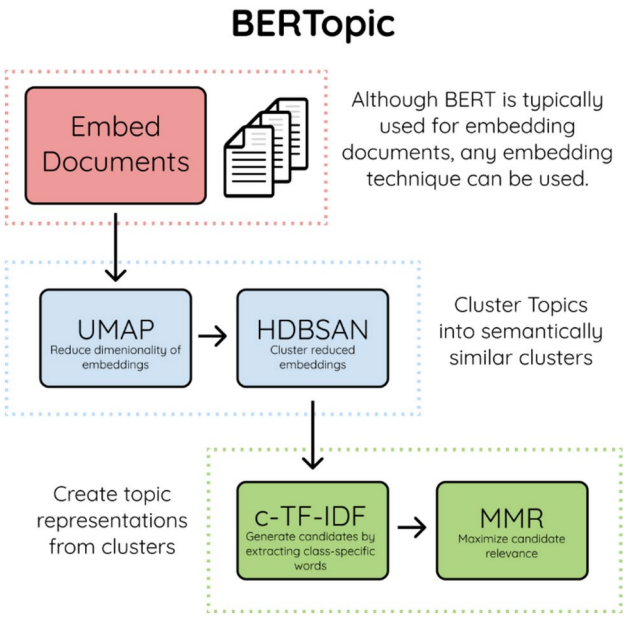

Fig. 2. The research process of study[29].

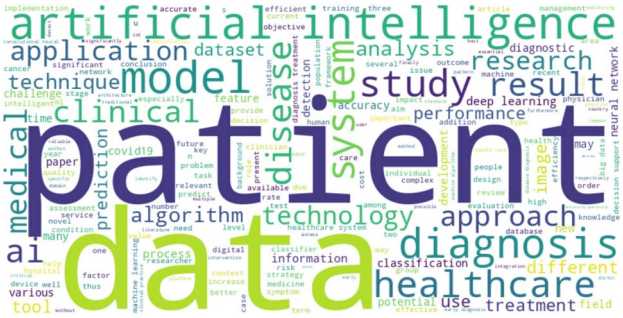

Fig. 3. Word cloud results for paper.

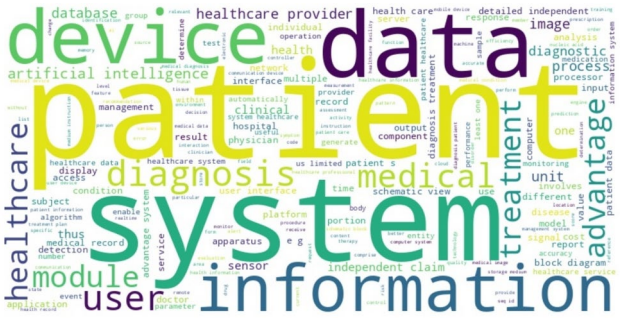

Fig. 4. Word cloud results for patent.

is then applied to refine keywords, balancing relevance and novelty for better topic distinction[46]. The equations are as follows.

$core_k^x = dx, N^k x$[53].

$d_{mreach-k}a, b = maxcore_k a, core_k b, da, b$.[53].

$W_{t,d} = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right)$[29].

It is calculated by dividing the number of documents in the corpus N by the total number of documents containing t, and then taking the logarithm; where $W_{t,d}$ refers to the weight of term t in document d; $tf_{t,d}$ is the term frequency of term t in document d, that is, the number of times term t appears in document d; N is the total number of documents in the document collection; df_t is the number of documents containing the term t, that is, the number of documents in which term t appears; $\log\left(\frac{N}{df_t}\right)$ is the calculation method of inverse document frequency, and the higher the value, the less common term t is in the entire document collection, and therefore the more important it is within a particular document[46]. This step not only identifies the core concepts of each topic but also provides an intuitive vocabulary list for understanding the uniqueness of each cluster. Words in a topic may be like each other and may be redundant for the explanation of the topic. In theory, this problem can be solved by applying the maximum marginal relevance to the first n words in topic[46]. This algorithm selects the most representative key words by balancing their relevance and novelty.

Results

This study uses text analysis based on the BERT model, which effectively overcomes the subjective judgment bias that may exist in traditional expert evaluation. Utilizing this model, information and patterns can be automatically extracted from large amounts of text data. In addition, BERT is highly competent at processing the deep semantic relationships of natural language and accurately understanding complex contexts and metaphors, which cannot be achieved by keyword frequency analysis alone[55]. Research results show that in health care, academic papers related to AI diagnosis can be traced back to 2001, while related patents began in 2003.

Word cloud

As shown in Fig. 3, the word cloud results of papers related to AI diagnosis in healthcare highlights several prominent themes, including “patient”, “clinical”, “healthcare”, “diagnosis”, “model”, “algorithm”, “data” and “artificial intelligence”. In particular, the prominence of the words “patient” and “clinical” indicates that the focus of the research is likely to be on issues about patients in clinical settings. In addition, the prominence of the word “clinical” also implies that AI technology has moved from theoretical research to clinical application and has had an impact on actual medical practice, especially in disease diagnosis.

According to the word cloud diagram of the Fig. 4 patent, the most prominent words are “patient”, “data”, “medical” and “information”, which indicate that patient data and medical information are the most frequently

discussed topics in the healthcare field. The words closely related to "diagnosis" and "treatment" indicate that data processing and information management during diagnosis and treatment are the focus of current research. For example, "artificial intelligence", "algorithm", and "sensor" appear in the word cloud, implying that medical devices and systems are becoming more intelligent, integrating advanced algorithms and sensors to optimise patient monitoring and health management. This may also indicate that, with the support of big data, personalised medicine and predictive health care may be future market investment hot spots.

Hierarchical clustering

Hierarchical clustering calculates the similarities between topics identified by the BERT model and builds these topics into a tree structure to reveal their hierarchical relationships and subtle connections[56]. As illustrated in Fig. 4, the paper shows the correlation and clustering of research topics in AI-related academic literature in the field of healthcare. Each branch of the dendrogram represents a different research topic, and the length of the branches and the connections between them reflect the similarities and differences between these topics. In this diagram, closely connected branches indicate that two topics have a high degree of correlation or similarity in keywords or research content.

According to Fig. 5, the most significant clustering occurs around specific medical subfields, such as COVID-19 research, cancer classification, cardiology, neurology, and the application of AI in healthcare. For example, topics like "coronavirus", "pandemic", and "COVID19" form a red cluster, showing that these topics are strongly related to each other. Similarly, cancer classification, cardiology, and neurodegenerative diseases are clustered together, revealing the trend of actively applying AI methods to the diagnosis and research of different disease types. For example, there might be a branch that closely connects a topic related to heart disease (such as "cardiology_cardiac_cardiologist") to other cardiovascular disease topics, showcasing these research areas share similar methods or data types when using AI technology. Furthermore, this hierarchical clustering can reveal large-scale trends in medical AI research. If the topics of "deep learning", "classification" and "image processing" are very closely related, it indicates that in the field of image processing, deep learning and classification algorithms are being actively researched and applied to the analysis of medical imagery.

The hierarchical clustering of patents reveals the relationship between patent topics related to AI in healthcare. The clusters reflect the key areas of current medical technology, including the continuous integration of cloud computing, big data, telemedicine, sensor monitoring, medical imaging, etc. The short horizontal lines at the top of the dendrogram represent highly similar topics that are grouped to form a cluster. In contrast, long horizontal lines indicate larger differences between subjects. For example, some closely related medical data processing topics may be clustered together, showing that they have commonalities in methods, types of data used, or research purposes. Topics that are relatively scattered in the figure, such as "quantum_gc_computing" and "biometric_sensor_monitoring", may point to more unique or specific research areas. The diversity of some clustering themes can also be observed in Fig. 6. For instance, green branch topics surrounding "cloud_computing_encryption," "dataset_database_processing," and "blockchain_database_computing" may represent research in cloud computing, data processing and blockchain technologies that are becoming increasingly important in medical data security and management. Furthermore, the red branches focus on topics such as "telemedicine", "medical monitoring services" and "diagnostic data", which may reflect research trends in telemedicine services and diagnostic technologies that are becoming increasingly critical in modern healthcare. The purple, yellow and cyan clusters involve sensor technology, vital sign monitoring and respirators, showing the growing application of sensors and biomarkers in medical diagnosis and monitoring. The grey clusters focus

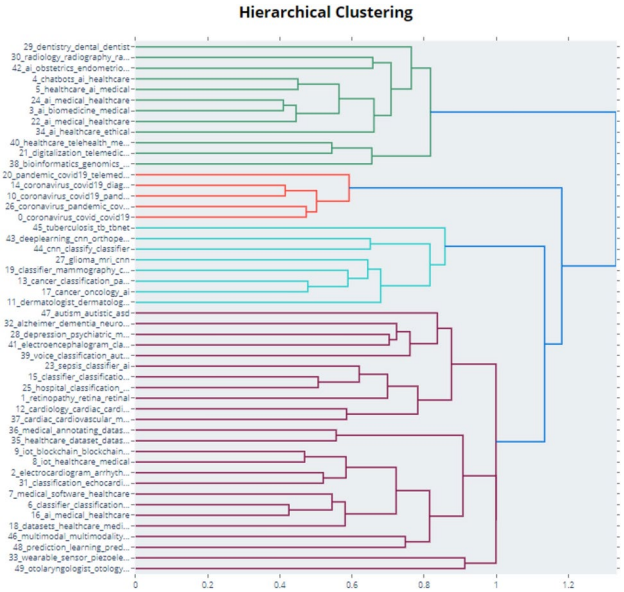

Fig. 5. Hierarchical clustering results for paper.

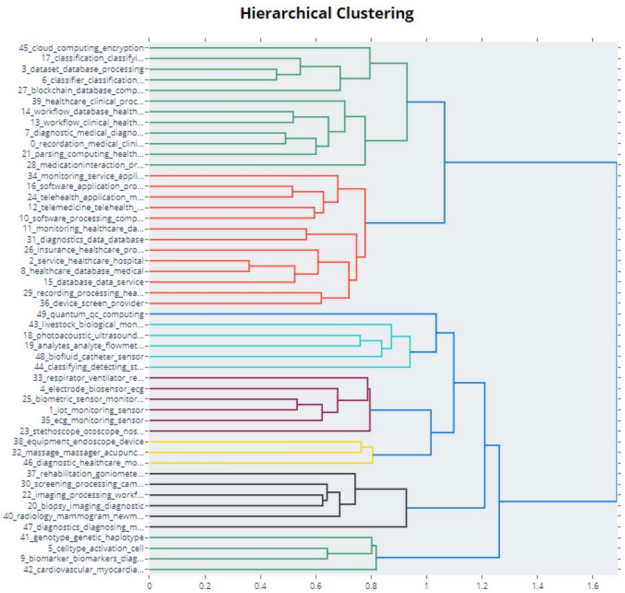

Fig. 6. Hierarchical clustering results for patent.

on image processing and biomedicine, indicating that AI research in medical image analysis and radiology is developing rapidly. These clusters reveal not only the widespread application of cloud computing, big data, telemedicine, sensor monitoring and medical image analysis in the medical field, but also the growth of data processing and cloud computing, the rise of telemedicine and monitoring technology, the widespread application of sensor technology and the rapid development of AI applications in medical imaging. These trends are driving the realization of smarter, more efficient medical services and personalized medicine.

Evolution path analysis

By comparing the trends and life cycles of AI in the healthcare field in scientific papers and patents, this study can observe the development trajectories of different topics and their corresponding market and scientific research dynamics. The topic trends in scientific papers are relatively stable and continuously rising, which reveals the continuity and depth of scientific research. The trends in patent charts are more volatile, which is closely related to the commercialisation process of technology development and the life cycle of patents. According to Fig. 7, topics such as diabetic eye disease and cardiac imaging AI show a continued growth trend in scientific papers, reflecting the academic community's long-term and in-depth research in these fields. Meanwhile, pandemic topics spike during specific periods, such as the COVID-19 pandemic, illustrating the concentrated surge in emergency response research. In terms of patent trends as shown in Fig. 8, topics such as wearable health monitoring and medical imaging devices show fluctuating growth, which may be the result of the interaction of technological advancements and market demand. Fluctuations in topics such as AI data processing, patient medical databases and electronic health monitoring within a given year reflect the cyclical nature of the patent filing and approval process. In addition, diagnostic radiology shows strong growth in both fields, indicating that its AI applications have received widespread attention and investment from scientific research and industry. Finally, the growth in patents for telemedicine application development suggests that technology companies may be actively developing new telemedicine technologies, driven in part by the increased demand for remote services driven by the COVID-19 pandemic.

Discussion

This study expands the application of AI in healthcare technology life cycle theory and proposes a hybrid research framework using both papers and patents. From a theoretical perspective, this study clarifies the relationship and difference between paper and patent in AI in healthcare, and provides significant direction and guidance for subsequent theoretical research. Word cloud analysis intuitively reveals that scientific literature and patent documents have different focuses and trends regarding the rise of AI in healthcare. The word cloud of scientific papers indicates that research papers focus mainly on the development of AI theory, modelling of clinical applications, algorithm design and predictive applications in diagnosis. In the patent word cloud, words such as "patient", "device", "data", "medical" and "information" are highlighted, suggesting that patent documents focus more on the practical application of artificial intelligence technology, such as medical equipment, data processing and medical information systems. The word cloud used in patent documents highlights the specific implementation of the technology, the solution applied to patients, and the management of medical data. Compared to the theoretical and predictive research of scientific papers, patents focus more on the innovation, practicality and commercialisation potential of the product.

In the hierarchical clustering results, the relationships and clustering patterns between the respective research topics can be observed. Hierarchical clustering of patents may place more emphasis on technological

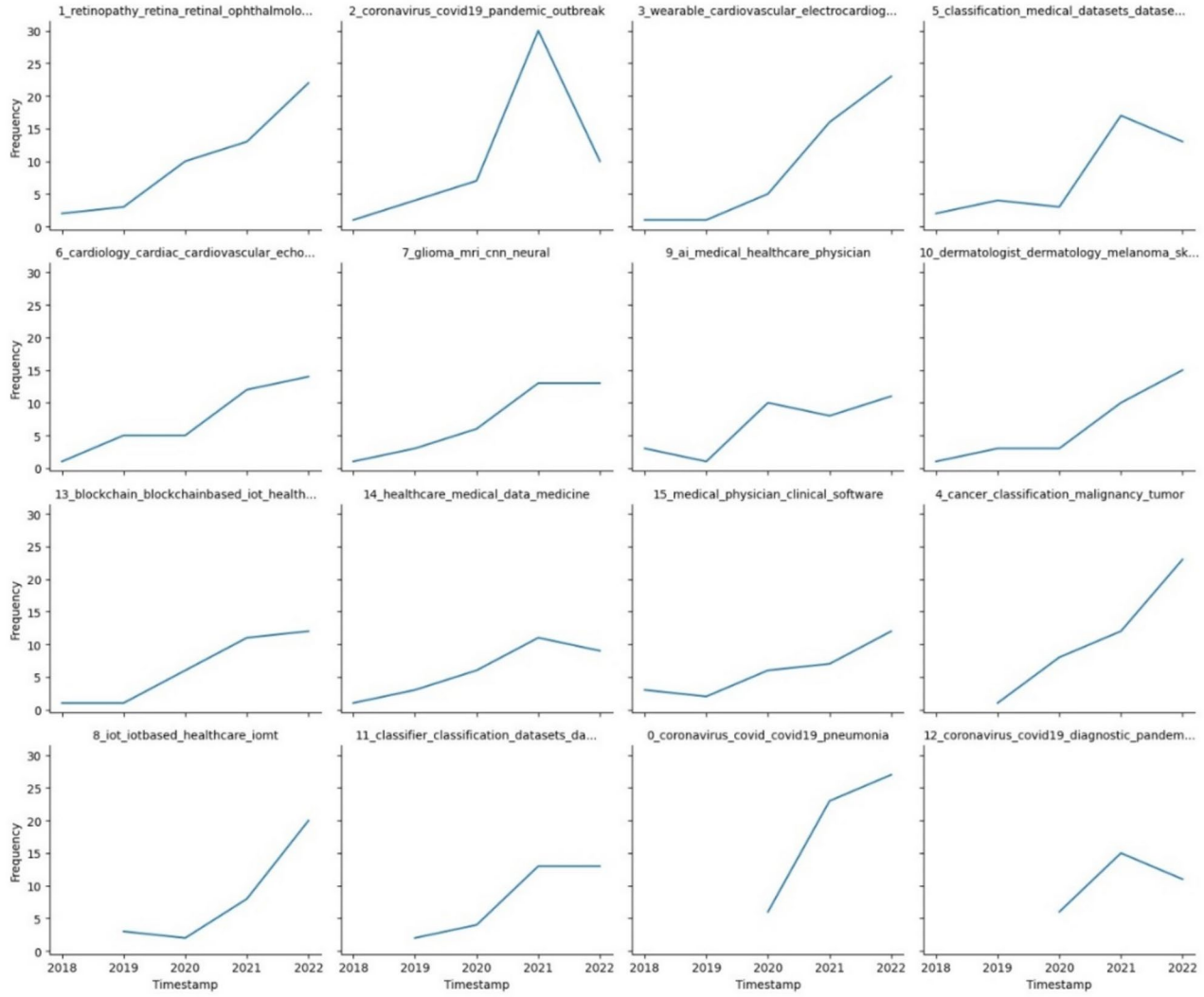

Fig. 7. Trend analysis (topic over time) for paper.

implementation and application innovation. For example, the technology topics that come together may involve the IoT, data encryption, medical devices, etc. Within these clusters, we can find trends that focus on technical details, product development and application deployment. In comparison, the clustering of academic papers may be more focused on exploring new theoretical concepts, developing advanced algorithms and evaluating the potential of AI technology in medical diagnosis. Clusters in papers may include research in various biomedical fields, such as biomarker discovery, genetic analysis of disease and evaluation of novel diagnostic tools. Therefore, the clustering of patents will revolve around technologies and products that can be commercialised, while the clustering of academic papers may include broader research areas aimed at promoting the understanding and development of knowledge within the scientific community.

Analysing the trends and technology life cycle of BERT results reveal a steady increase in the frequency of topics such as diabetes, cardiac monitoring, and clinical data analysis in scientific papers. Research response rates of research are also relatively rapid during global health crises and other situations requiring urgent research. Patent literature analysis shows that growing market demand has driven innovation in technologies such as health care data analysis and telemedicine application development. Patent trends are more volatile, reflecting the cyclical changes in technology commercialisation. Technologies such as wearable health monitoring, the IoT and medical imaging equipment show fluctuating growth, indicating the interaction between market demand and technological progress.

The transformation of healthcare from traditional methods to new approaches based on data and AI, while offering numerous improvements and conveniences, also presents potential problems and risks, particularly regarding data security. Challenges around sexuality and privacy are particularly acute. Firstly, as healthcare systems increasingly rely on electronic data recording and storage, sensitive data ranging from patients' personal health information to treatment histories are stored on digital platforms. Access to this information by unauthorized third parties may result in serious privacy breaches and other security risks. Secondly, privacy issues are also extremely challenging. In the use of new medical technologies, especially those involving big data

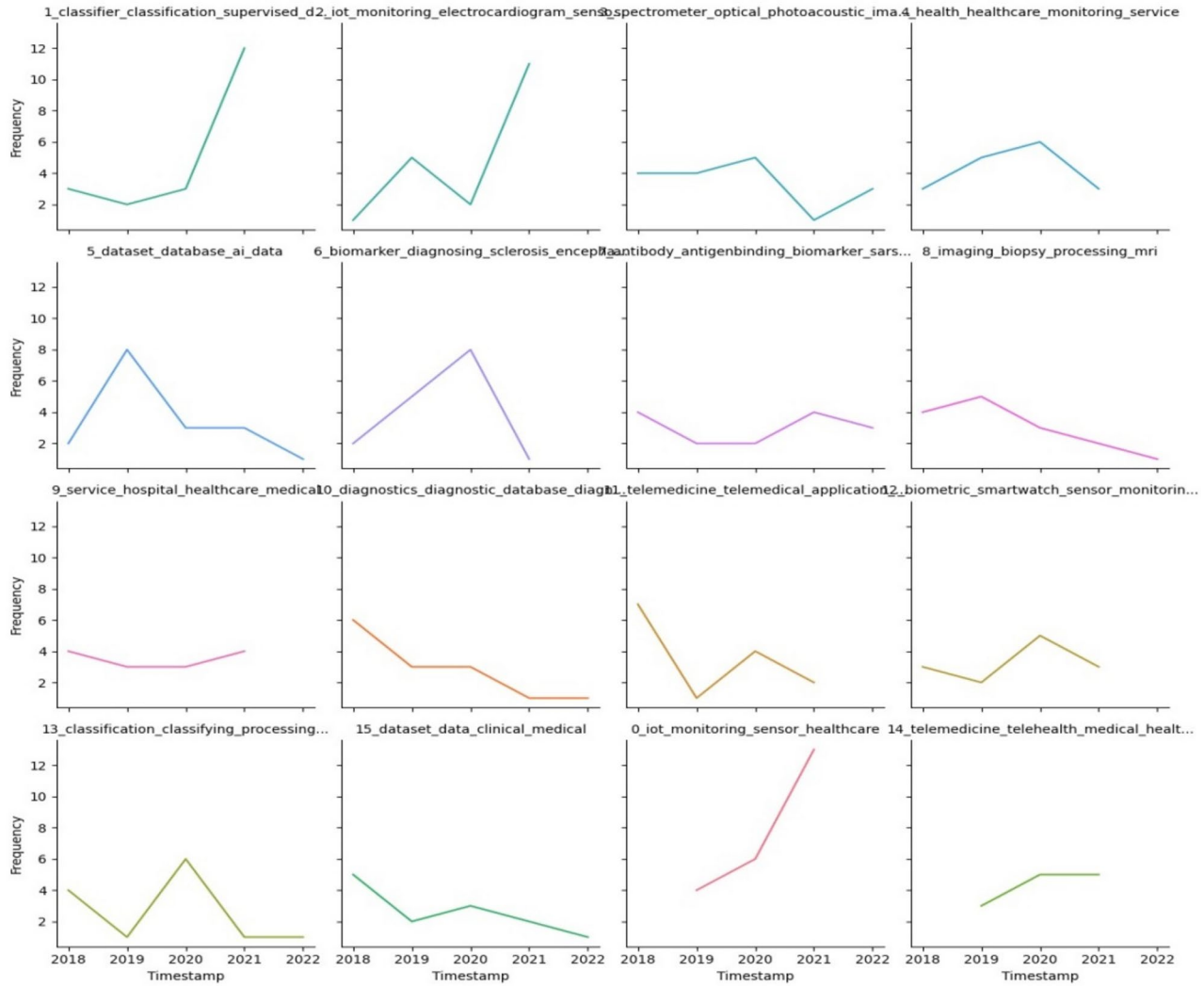

Fig. 8. Trend analysis (topic over time) for patent.

and AI, ensuring that patients' privacy rights are not violated is a key concern. Apart from these, legal and policy frameworks are needed to regulate the application of these technologies and protect patient privacy and data security. Data security and AI ethics issues are currently vital topics and potential challenges.

Conclusions

This research uses BERT-based text analysis methods to explore future trends in emerging AI applications within healthcare, drawing on both scientific papers and patents. The research shows that, in the past few years, AI's application in healthcare has made significant leaps, gradually expanding from focused theoretical research to practical clinical applications, and demonstrating continued growth in patent activity.

By analysing keyword word cloud, hierarchical clustering, topic clustering and evolution analysis of academic papers and patents, this study gained insights into the evolution path of AI technology in the healthcare industry and its potential impact on future healthcare practice. In addition, word cloud analysis reveals the research focus of AI applications in the medical field and their interconnections and clustering patterns. In particular, the frequent occurrence of keywords, such as "AI" and "clinical" in scientific papers points to the focus of research on applying AI technologies to clinical settings to improve the quality of medical services. In patent documents, the frequent occurrence of words such as "data" and "patient" suggests a focus on innovations in technology implementation and data processing. Through trend analysis of different time series, the dynamic changes and technology evolution path of AI in healthcare are derived. From rapidly rising research interest to a sign of technological maturity, the development of medical AI technology has shown obvious cyclicality and reactivity. Especially during global health events such as the COVID-19 pandemic, research and technology development activities increased significantly. From cardiology to diabetic eye disease, from telemedicine to intelligent diagnostic systems, medical AI technology is gradually becoming a core force in promoting the modernisation of healthcare. In the process of transforming from traditional medical methods to new medical treatments, some

potential issues and risks are also being considered, such as data security and privacy that need to be addressed. This can be a great potential direction for future research.

Above all, the main contribution of this paper is the introduction of the BERT model in the field of AI in healthcare for the first time. By applying the pre-trained BERT model for text mining and deep learning, the topic modelling and trend analysis of AI in healthcare avoids the shortcomings of traditional models that rely on expert judgment, providing a more accurate and objective research method. The innovation and novelty also lie in the combined use of two data sources: paper and patent. Academic papers mainly focus on theoretical innovations and research trends, while patents are key resources for understanding the commercialisation and practical application of emerging technologies. Integrating these two data sources allows this study to simultaneously capture the conceptual progress and practical applications of AI in healthcare, identifying market trends, commercialisation paths, and promising innovations that have not been widely studied but are under active development.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 30 October 2024; Accepted: 21 February 2025
Published online: 04 March 2025

References

1. McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AIMag* **27**, 12–12 (2006).
2. Wang, Y. J., Choo, W. C. & Ng, K. Y. Review and bibliometric analysis of AI-driven advancements in healthcare. *APJMBB* https://doi.org/10.35118/apjmbb.2024.032.2.10 (2024).
3. Cowls, J., Tsamados, A., Taddeo, M. & Floridi, L. A definition, benchmark and database of AI for social good initiatives. *Nat. Mach. Intell.* **3**, 111–115 (2021).
4. Wang, J. F., Zhang, Z. X., Feng, L. J., Lin, K. Y. & Liu, P. Development of technology opportunity analysis based on technology landscape by extending technology elements with BERT and TRIZ. *Technol. Forecast. Soc. Change* **191**, 122481 (2023).
5. Sharma, C. et al. Predicting trends and research patterns of smart cities: A semi-automatic review using latent Dirichlet allocation (LDA). *IEEE Access* **10**, 121080–121095 (2022).
6. Alowais, S. A. et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
7. Wang, Y. J. *Technology Adoption in Health Care System: Cross-Cultural Investigation of AI Adoption in China and UK (Master Dissertation)* (2020).
8. Alwabel, A. S. A. & Zeng, X.-J. Data-driven modeling of technology acceptance: A machine learning perspective. *Expert Syst. Appl.* **185**, 115584 (2021).
9. Dahlke, J. et al. Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. *Res. Policy* **53**, 104917 (2024).
10. Madan, R. & Ashok, M. AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Gov. Inf. Q.* **40**, 101774 (2023).
11. Thouin, M. F., Hoffman, J. J. & Ford, E. W. The effect of information technology investment on firm-level performance in the health care industry. *Health Care Manag. Rev.* **33**, 60–68 (2008).
12. Sabherwal, R. & Jeyaraj, A. Information technology impacts on firm performance: An extension of Kohli and Devaraj (2003). *MISQ* **39**, 809–836 (2015).
13. Dong, J. Q., Karhade, P. P., Rai, A. & Xu, S. X. How firms make information technology investment decisions: Toward a behavioral agency theory. *J. Manag. Inf. Syst.* **38**, 29–58 (2021).
14. Li, X., Xie, Q., Daim, T. & Huang, L. Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology. *Technol. Forecast. Soc. Change* **146**, 432–449 (2019).
15. Benchimol, J., Kazinnik, S. & Saadon, Y. Text mining methodologies with R: An application to central bank texts. *Mach. Learn. Appl.* **8**, 100286 (2022).
16. Nitiéma, P. Artificial intelligence in medicine: Text mining of health care workers' opinions. *J. Med. Internet Res.* **25**, e41138 (2023).
17. Lee, C. Y. A review of data analytics in technological forecasting. *Technol. Forecast. Soc. Change* **166**, 120646 (2021).
18. Erzurumlu, S. S. & Pachamanova, D. Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. *Technol. Forecast. Soc. Change* **156**, 120041 (2020).
19. Dwivedi, Y. K. et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **57**, 101994 (2021).
20. Khan, M. et al. Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. *Expert Syst. Appl.* **185**, 115695 (2021).
21. Zahlan, A., Ranjan, R. P. & Hayes, D. Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research. *Technol. Soc.* **74**, 102321 (2023).
22. Zhang, K. & Aslan, A. B. AI technologies for education: Recent research & future directions. *Comput. Educ. Artif. Intell.* **2**, 100025 (2021).
23. Cho, I. & Ju, Y. Text mining method to identify artificial intelligence technologies for the semiconductor industry in Korea. *World Patent Inf.* **74**, 102212 (2023).
24. HaCohen-Kerner, Y., Miller, D. & Yigal, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One* **15**, e0232525 (2020).
25. Subakti, A., Murfi, H. & Hariadi, N. The performance of BERT as data representation of text clustering. *J. Big Data* **9**, 15 (2022).
26. Zhou, X., Huang, L., Zhang, Y. & Yu, M. A hybrid approach to detecting technological recombination based on text mining and patent network analysis. *Scientometrics* **121**, 699–737 (2019).
27. Adlung, L., Cohen, Y., Mor, U. & Elinav, E. Machine learning in clinical decision making. *Med* **2**, 642–665 (2021).
28. Gui, M. & Xu, X. Technology forecasting using deep learning neural network: Taking the case of robotics. *IEEE Access* **9**, 53306–53316 (2021).
29. Gupta, T., Zaki, M., Krishnan, N. M. A. & Mausam,. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput. Mater.* **8**, 1–11 (2022).
30. Rong, G., Mendez, A., Bou Assi, E., Zhao, B. & Sawan, M. Artificial intelligence in healthcare: Review and prediction case studies. *Engineering* **6**, 291–301 (2020).

31. Kaul, V., Enslin, S. & Gross, S. A. History of artificial intelligence in medicine. *Gastrointest. Endosc.* **92**, 807–812 (2020).
32. Vo, N. N. Y., He, X., Liu, S. & Xu, G. Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decis. Support Syst.* **124**, 113097 (2019).
33. Wang, D., Su, J. & Yu, H. Feature extraction and analysis of natural language processing for deep learning english language. *IEEE Access* **8**, 46335–46345 (2020).
34. Sharma, C., Sakhuja, S. & Nijjer, S. Recent trends of green human resource management: Text mining and network analysis. *Environ. Sci. Pollut. Res.* **29**, 84916–84935 (2022).
35. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).
36. Zhao, F. et al. Multi-layer features ablation of BERT model and its application in stock trend prediction. *Expert Syst. Appl.* **207**, 117958 (2022).
37. Belwal, R. C., Rai, S. & Gupta, A. Text summarization using topic-based vector space model and semantic measure. *Inf. Process. Manag.* **58**, 102536 (2021).
38. Rashid, J., Shah, S. M. A. & Irtaza, A. Fuzzy topic modeling approach for text mining over short text. *Inf. Process. Manag.* **56**, 102060 (2019).
39. Passalis, N. & Tefas, A. Learning bag-of-embedded-words representations for textual information retrieval. *Pattern Recognit.* **81**, 254–267 (2018).
40. Antons, D., Grünwald, E., Cichy, P. & Salge, T. O. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R & D Manag.* **50**, 329–351 (2020).
41. Lee, J. Y., Ahn, S. & Kim, D. Deep learning-based prediction of future growth potential of technologies. *PLOS One* **16**, e0252753 (2021).
42. Ohri, K. & Kumar, M. Review on self-supervised image recognition using deep neural networks. *Knowl.-Based Syst.* **224**, 107090 (2021).
43. Le, E. P. V., Wang, Y., Huang, Y., Hickman, S. & Gilbert, F. J. Artificial intelligence in breast imaging. *Clin. Radiol.* **74**, 357–366 (2019).
44. Sampaio, P. G. V. et al. Photovoltaic technologies: Mapping from patent analysis. *Renew. Sustain. Energy Rev.* **93**, 215–224 (2018).
45. Gupta, P., Ding, B., Guan, C. & Ding, D. Generative AI: A systematic review using topic modelling techniques. *Data Inf. Manag.* **8**, 100066 (2024).
46. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Preprint at https://doi.org/10.48550/ARXIV.2203.05794 (2022).
47. Zhu, Y. & Zhang, J. Technology life cycle embedded technology development path analysis method. *Procedia Comput. Sci.* **202**, 289–294 (2022).
48. Yang, Y. et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep.* **36**, 109442 (2021).
49. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
50. Hozumi, Y., Wang, R., Yin, C. & Wei, G.-W. UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Comput. Biol. Med.* **131**, 104264 (2021).
51. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/ARXIV.1802.03426 (2018).
52. Vijayan, D. & Aziz, I. Adaptive hierarchical density-based spatial clustering algorithm for streaming applications. *Telecom* **4**, 1–14 (2022).
53. Huang, J. et al. Fire risk assessment and warning based on hierarchical density-based spatial clustering algorithm and grey relational analysis. *Math. Probl. Eng.* **2022**, 1–8 (2022).
54. Sood, P., Sharma, C., Nijjer, S. & Sakhuja, S. Review the role of artificial intelligence in detecting and preventing financial fraud using natural language processing. *Int. J. Syst. Assur. Eng. Manag.* **14**, 2120–2135 (2023).
55. Wei, C., Chaoran, L., Chuanyun, L., Lingkai, K. & Zaoli, Y. Tracing the evolution of 3-D printing technology in China using LDA-based patent abstract mining. *IEEE Trans. Eng. Manag.* **69**, 1135–1145 (2022).
56. George, L. & Sumathy, P. An integrated clustering and BERT framework for improved topic modeling. *Int. J. Inf. Tecnol.* **15**, 2187–2195 (2023).

Author contributions

Y.J.W. conceived and designed the analysis, contributed data or analytical tools, performed the analysis, and wrote the paper. W.C.C.conceived and designed the study, supervised and provided methodological guidance. K.Y.N was involved in data organization and combing, and provided guidance on data analysis and interpretation. B.R. assisted in data interpretation, analyzed and interpreted data, and optimized models. P.W.W. designed the analysis, and critically revised the manuscript for important intellectual content. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.W.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.