Hindawi
Scientific Programming
Volume 2022, Article ID 3212370, 12 pages
https://doi.org/10.1155/2022/3212370

Review Article

A Review on the Application of Knowledge Graph Technology in the Medical Field

Jia Qu

Information Center, Hebei Petroleum University of Technology, Chengde 067000, China

Correspondence should be addressed to Jia Qu; 15696966@qq.com

Received 16 April 2022; Accepted 7 July 2022; Published 20 July 2022

Academic Editor: Jianping Gou

With the continuous development of Internet technology, knowledge graph construction has received increasing attention. Extracting useful medical knowledge from massive data is the key to analyzing big medical data. The knowledge graph is a semantic network that reveals relationships between entities. Medicine is one of the widely used fields of knowledge graphs, and the construction of a medical knowledge graph is also a research hotspot in artificial intelligence. Knowledge graph technology has broad application prospects in the field. First, this study comprehensively analyzes the structure and construction technology of the medical knowledge graph according to the characteristics of big data in the medical field, such as strong professionalism and complex structure. Second, this study summarizes the key technologies and research progress of the four modules of the medical knowledge graph: knowledge representation, knowledge extraction, knowledge fusion, and knowledge reasoning. Finally, with the major challenges and key problems of the current medical knowledge graph construction technology, its development prospects are prospects.

1. Introduction

Since the semantic web was proposed by Berners-Lee, the father of the World Wide Web, in 1998, people have continued to express and revise their understanding of the objective world on electronic carriers such as the Internet, forming a process of conceptual standardization; at the same time, with the rapid growth of the scale of linked open data, more and more knowledge metadata are scattered on the Internet. The knowledge graph is a way of knowledge representation and management under the background of big data, which emphasizes the ability of semantic retrieval. With the vigorous development of artificial intelligence in recent years, the key problems involved in the knowledge graph, such as knowledge extraction, representation, fusion, reasoning, question, and answer, have been solved and broken through to a certain extent. Knowledge graph has become a new hot spot in knowledge service. It has been widely concerned by scholars and industry at home and abroad.

The knowledge graph is a concept proposed by Google in 2012, essentially the semantic web's knowledge base. The knowledge graph consists of nodes and edges. The nodes represent entities, and the edges represent the relationship between entities. This is the most intuitive and understandable framework for knowledge representation and reasoning and lays the foundation for the third generation of artificial intelligence research [1]. With the continuous development and continuous transformation of information technology and Internet technology, human beings have successively experienced the "Web 1.0" era characterized by document interconnection and the "Web 2.0" era characterized by data interconnection—the new "Web3.0" era of knowledge interconnection [2]. Knowledge interconnection aims to build a World Wide Web that both humans and machines can understand and make the network more intelligent. However, the multisource heterogeneity and loose organizational structure of the content on the World Wide Web bring great challenges to the knowledge interconnection in the big data environment [2]. Therefore, people need to explore the knowledge interconnection method that not only conforms to the constantly changing laws of network information resources but also meets the cognitive

needs of users according to the principle of organizational knowledge in the big data environment [3], to make it more profoundly display the overall and interrelated knowledge [4]. The knowledge graph is a way of knowledge representation and management generated under such a background. It is the foundation and bridge for realizing intelligent semantic retrieval and lays a solid foundation for knowledge interconnection on the World Wide Web [5].

The knowledge graph is the frontier research problem of intelligent big data. It conforms to the development of the information age with unique technical advantages, such as incremental data pattern design, good data integration, existing RDF, OWL, and other standards support, semantic search and knowledge reasoning ability, and so on. With the development of regional health information and medical information systems in the medical field, much medical data has been accumulated. Promoting medical intellectualization is the key problem in extracting information from these data and managing, sharing, and applying them. It is the basis of medical knowledge retrieval, clinical diagnosis, medical quality management, electronic medical records, and intelligent processing of health files.

At present, medicine is one of the most widely used vertical fields of the knowledge graph, and it is also a hot spot in the field of artificial intelligence at home and abroad. It has a good development prospect in intelligent medicine, such as disease risk assessment, intellectual assistant diagnosis and treatment, medical quality control, and medical knowledge question and answer [5]. At present, many companies have constructed their knowledge graph. The application of medical knowledge graph has also entered people's attention in the past two years, such as IBM's Watson Health, Ali Health's medical think tank, and Sogou's AI medical knowledge graph APGC. In the medical field, typical medical knowledge graphs include SNOMED-CT, IBM's Waston Health, and traditional Chinese medicine knowledge graphs such as Shanghai Shuguang Hospital [6]. With regional health informatization and medical information technology development, many medical data has accumulated. Extracting and applying information from these data is the key to promoting intelligent medical treatment [7]. It is also the basis of medical knowledge retrieval, auxiliary diagnosis and treatment, medical quality control, electronic medical records, and intelligent health management. It is significant to improve doctors' diagnosis and treatment levels and lighten doctors' burden.

The main contributions of this study can be summarized as follows:

(1) This paper comprehensively sorts out the key technologies and applications of medical knowledge graph construction and reviews various public data sets, difficulties in dealing with medical problems, and existing solutions.

(2) By reading this article, you can understand the development status, future development direction, and challenges of the medical knowledge graph, which is convenient for researchers to refer to and compare and accelerate the research and clinical application in the field of the medical knowledge graph.

(3) This paper conducts a comprehensive investigation and in-depth analysis of the key technologies of medical knowledge graph construction and the current research and application development status. It summarizes the important challenges and key issues faced by the construction of medical knowledge graphs.

2. Methods

2.1. Construction of Medical Knowledge Graph. This paper summarizes the medical knowledge graph construction technology into four parts: medical knowledge representation, extraction, fusion, and reasoning. By extracting the constituent elements of the knowledge graph such as entities, relationships, and attributes from a large amount of structured or unstructured medical data, select a reasonable and efficient way to store them in the knowledge base. Medical knowledge fusion disambiguates and links the content of the medical knowledge base, enhances the internal logic and expressive ability of the knowledge base, and updates old knowledge or supplements new knowledge for the medical knowledge graph through manual or automatic means; with the help of knowledge reasoning, it is inferred. In the absence of facts, disease diagnosis and treatment can be automatically completed; quality assessment is an essential means of ensuring data and improving the credibility and accuracy of the medical knowledge map.

This paper comprehensively combs the key technologies and applications of the construction of medical knowledge graph. It summarizes public data sets, specific difficulties in dealing with medical problems, and existing solutions. By reading this article, we can understand the development status, future development direction, and challenges of the medical knowledge graph; facilitate the reference and comparison of medical knowledge graph researchers; and speed up the research and clinical application in the medical knowledge graph. This paper mainly describes the process of constructing medical knowledge graph, and the framework of the medical knowledge graph is shown in Figure 1.

2.2. Medical Knowledge Representation. Knowledge representation is a set of conventions made to describe the world and is a process of knowledge symbolization, formalization, and modeling [8]. It mainly studies the methods of computer knowledge storage, and its representation affects the efficiency of knowledge acquisition, storage, and application of the system. However, with the characteristics of various types of medical data, different storage methods, formats, and standards of electronic medical records and often involving cross-fields, there are some differences in knowledge representation between the medical field and other fields. At the same time, it also challenges the knowledge representation in the medical area.

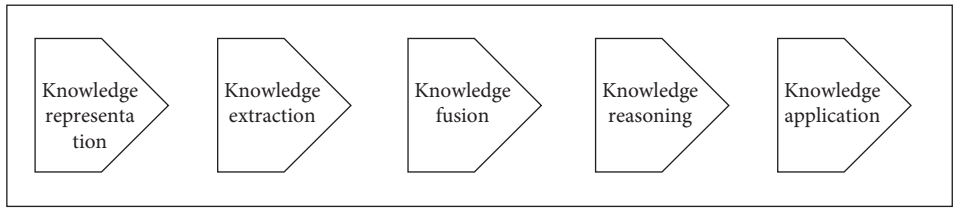

Figure 1: Medical knowledge graph framework.

The knowledge representation methods used in early medical knowledge bases include predicate logic representation, production representation, frame representation, and semantic web representation, such as SNOMED CT, early MYCIN system [8], Escherichia coli database EcoCyc, and so on [9]. With the growth of knowledge and the complexity of relationships in the knowledge graph, these methods are no longer the primary knowledge representation method due to their limited representation ability and lack of flexibility but more as an auxiliary or supplement to medical knowledge representation.

Ontology representation represents knowledge in the form of a network. Two associated nodes (entities) are represented by triples (entity 1, relationship, entity 2) and gradually recognized after the knowledge graph was proposed. It draws on the semantic web notation but has some differences. Ontology focuses on the inherent characteristics of entities, which is more focused and deeper than the latter, so it also has greater development potential. Ontology description languages are diverse, mainly RDF and RDFS, DAML, OWL, and so on [9]. Using ontologies to represent medical terms can improve data integration capabilities; build robust and interoperable medical information systems; meet the needs of reusing, sharing, and transmitting medical data; and provide statistical aggregation based on different semantic standards. The construction of medical domain ontology requires an in-depth analysis of the structure and concepts of medical terms to express obscure and even cross-language medical knowledge effectively. The current medical knowledge ontology database includes the medical concept knowledge database LinkBase [10], TAMBIS ontology database [11], and so on.

The representative models in knowledge representation learning are SE (structure embedding) [12], SLM (single-layer model) [13], LFM (latent factor model) [14], translation model based on TransE [15], and so on. These models consider the cooperation and computational overhead between entities, use vectors to represent entities, carry out the corresponding matrix transformation of the vectors or relations representing entities, and put forward an evaluation function to measure the correlation between entities. And provide an important reference for later knowledge completion and reasoning. Kleyko et al. [16] proved that the distributed representation method could represent medical images for classification with the same accuracy as the best classical method. Henriksson et al. [17] compared multiple knowledge representation methods to represent four types of records in EHR: diagnostic records, drug use records, treatment methods, and disease course records. Knowledge representation learning undoubtedly opens up a new idea for the knowledge representation of medical knowledge graphs.

2.3. Knowledge Extraction. The construction of medical knowledge graph is mainly to extract entities, relationships, and attributes manually or automatically from unstructured data. Manual extraction is through experts to collect and sort out relevant information according to certain rules to extract knowledge. The manually constructed medical knowledge base includes clinical medical knowledge [17], SNOMED-CT, ICD-10 [29], and so on. Automatic extraction automatically uses machine learning, artificial intelligence, data mining, and other information extraction technology to extract basic knowledge graph elements from data sources. A typical example of automatic construction of medical knowledge base is the integrated medical language system UMLS [18]. The cost of manual extraction is too high, and automatic knowledge extraction is the current key research direction and the future trend of constructing knowledge graphs. This section automatically extracts knowledge and information from data sources, including entity, relationship, and attribute extraction. The common knowledge base of the above common medical ontology is shown in Table 1.

2.3.1. Entity Extraction. The entity is an essential element in the medical knowledge graph. The accuracy and recall rate of entity extraction will directly affect the quality of the knowledge base, so entity extraction is the key research direction of medical knowledge graph technology. The purpose of identifying biomedical entities in the text is to further extract relationships and other information by identifying key concepts and to express the identified concepts in a standardized form. Entity extraction in the medical field is to extract specific types of named entities from medical data sources, and the extraction methods of medical entities are classified into the following three categories:

(1) Methods based on medical dictionaries and rules

This method is challenging to generate dictionaries by manually defining rules and pattern matching or extracting medical entities from the corpus using existing medical dictionaries. First, at present, there is no complete dictionary covering all types of biological named entities, so a simple text matching algorithm is not enough to deal with entity recognition; second, the meaning of the same word or phrase can refer to different objects according to the

Table 1: Common knowledge bases of medical ontology.

Name	Data type	Amount of data
SNOMED-CT	Standard of clinical medical terminology	146,217
UMLS	Unified Medical Language System	Concept: 3,000,000 Name: 1,200,000
CMeKG	Chinese medical knowledge graph	Diseases: more than 10,000 kinds Drugs: nearly 20,000 Symptoms: more than 10,000 kinds Diagnosis and treatment technology and equipment: 3,000 kinds Concepts, relationships, and attributes: 1,560,000

change of context; third, many biological or pharmaceutical entities have multiple names at the same time (e.g., PTEN and MMAC1 refer to the same gene). Therefore, medical dictionaries and rules are widely used only in the earliest days. HAN et al. [19] identify medical information in electronic medical records by defining semantic patterns and syntax. Wu et al. [20] obtained good experimental results using two medical dictionaries, CHV [21] and SNOMEDCT. Although this method can achieve high accuracy, it cannot completely solve the above problems and rely too much on dictionaries and rules written by experts to adapt to the continuous emergence of words in the medical field.

(2) Machine learning method based on medical data source and mathematical model

In this method, entity recognition is carried out using statistics and machine learning methods, combined with the training model of medical data sources. The most representative tagging corpus in English medical entity extraction is the English electronic medical record tagging corpus published by i2b2 2010 [22]. In addition, NCBI [23] corpus provides English medical entity tagging data. Researchers try to apply machine learning and statistical algorithms to entity extraction, train the model based on the characteristics of medical data, and then identify entities. Standard methods include support vector machine, artificial divine meridian network, hidden Markov model, conditional random field, and so on. Kazama et al. [24] use a support vector machine model for biomedical named entity recognition. To improve the training effect, word cache and unsupervised training are introduced. The experimental results show that the accuracy of this method in the GENIA medical data set is higher than that of the benchmark algorithm and can be efficiently applied to a large-scale knowledge base. Zhou et al. [25] proposed a maximum entropy algorithm as a hybrid machine learning algorithm and rule dictionary-based extraction algorithm. The experiment is carried out on the Medline data set, and the accuracy and recall rate of the experiment is more than 70%. The entity extraction method based on machine learning is faced with the problems of intermingled data quality and low specialization of manual tagging when used in the medical field. There is special research on reducing the dependence on data tagging. Its principle is continuously using massive unlabeled data to improve the model's performance. They are learning from small samples to form an interactive learning process to improve the accuracy of entity extraction.

(3) Deep learning method

Deep learning is a new field of machine learning that aims to establish and simulate the human brain for analytical learning [26]. It imitates the mechanisms of the human brain to interpret data, such as images, sounds, and texts. In recent years, it has been widely used in entity extraction. BiLSTM-CRF is the most mainstream deep learning model in medical entity extraction. Jagannatha and Hong [27] compared the effect of BiLSTM-CRF and other machine learning models in entity extraction of electronic medical records through experiments. The experimental results show that BiLSTM-CRF effectively improves the accuracy of results.

Most knowledge learning and deep learning methods need to collect a large amount of corpus or rely too much on the tagging of experts. Zhang et al. [28] proposed using the similarities and differences in the natural language representation of the labeled entity triple to encode the data distribution within various medical entity-relationship classes. Then, the unlabeled relational entity triples are found from the point of view of generating the model. This method reduces the overdependence of the traditional discriminant model on external resources and does not rely on the differences between medical entity relations. The experimental results show that the algorithm not only can generate entity triples belonging to a specific medical relationship with 92.91% support under the condition of limited external resources. The generated results have an accuracy of 77.17%, and 61.93% of the samples do not appear in the training data.

2.3.2. Relation Extraction. In this paper, medical entity relationship extraction is divided into two categories: the same type of hierarchical relationship extraction, such as gastrointestinal disease-chronic gastritis, and different types of relationship extraction, such as disease-symptom.

(1) Hierarchical relationship extraction of medical entities of the same type

The hierarchical relationship of the same medical entities is relatively simple, mainly "is-a" and "part-of" relationships. Because medicine has its rigorous discipline system and industry norms, such relationships are often carried out in medical dictionaries, encyclopedias, and information standards. Medical dictionaries or databases such as ICD-10 [29] and SNOMED focus on classifying and conceptualizing medical professional terms and restricted vocabulary. They are authoritative, cover a wide range, and are guaranteed quantity and quality. The medical industry widely uses them. Recognition is the preferred source for extracting entity relationships between hierarchies. According to the data format and open API interface provided by the specific medical dictionary and knowledge base, the hierarchical structure can be extracted from the crawler, regular expression, D2R mapping, and other technologies, and triples can be extracted to match and add the upper and lower relationship.

(2) Relation extraction of different types of medical entities

Semantic relationship recognition between different medical entities is roughly based on two different data sources. One is an encyclopedia or other structured data sources, such as MEDLINE, UMLS, and so on. The other is semistructured electronic medical records. The types of medical entities are relatively limited (mainly diseases, symptoms, treatments, drugs, etc.). The relationship types to be extracted are usually predefined between the two entities, and then the extraction task is transformed into a classification problem to deal with. There is no unified standard for predefining entity relations, which depends on the setting of the pattern graph, entity recognition, language sources, construction purpose, and application scenarios in constructing medical knowledge graph. For example, in the i2b2 2010, entity relations in electronic medical records are divided into three categories: medical problems, treatments, and examinations.

To reduce the transmission of errors to extraction in deep learning methods, in 2019, Eberts and Ulges [30] proposed a hybrid model including a transformer-based encoding layer, an LSTM entity detection module, and a reinforcement learning-based relation classification module. Experimental results show that the hybrid model performs better in relation and entity extraction than baseline methods. In 2019, Bansal et al. [31] proposed a new model SNERL (simultaneous neural entity-relation linker). A self-attention mechanism is first used to capture the contextual representations of each entity mentioned in the text; then these contextual representations are used to predict the mention-level entity distribution and the mention-pair-level relation distribution; finally, for each mention pair, the. These predicted probabilities are combined and merged at the document level to obtain the final likelihood of predicted relation triples. Experimental results show that the SNERL model achieves optimal performance on two biomedical datasets, CDT and CDR, and can significantly improve the overall recall of the system while avoiding cascading errors.

In response to the large span of medical relationships, in 2020, Nan et al. [32] proposed the LSR (latent structure refinement) model to construct a document-level graph end-to-end to reason about intersentence relationships. Through an iterative optimization strategy, the model can be dynamically constructed. The model achieves promising results on two document-level relation extraction data sets in the biomedical domain. The most commonly used data sets for medical entity relation extraction are seen in Table 2.

2.3.3. Attribute Extraction. Attribute extraction is for medical entities, such as the familiarity of drugs, including specifications, doses, indications, and so on. Entities can be completely delineated through attributes, such as metformin suitable for patients with type 2 diabetes because entities' attributes can be regarded as entities and attribute values. There is a name relationship between them to transform the attribute extraction problem into a relation extraction.

2.4. Knowledge Fusion. Due to the complex knowledge sources in medical databases, problems such as uneven knowledge quality, duplicate knowledge from different data sources, and ambiguous relationships between knowledge [33]. Under the same specification, the data are integrated, disambiguated, processed, reasoned, verified, updated, etc. of medical knowledge to judge the correctness of knowledge, and the roughness is extracted. Correct knowledge is organized into a knowledge base by aligning associations and merging calculations. Through the definition of knowledge fusion, it can be seen that knowledge fusion is based on knowledge extraction. Knowledge fusion research aims to eliminate uncertainty in knowledge understanding, discover the true value of knowledge, and expand the correct knowledge update into the knowledge base [34]. The key technologies of medical knowledge fusion include entity alignment technology, entity linking technology, and relationship deduction technology. Among them, entity alignment technology is used to eliminate the heterogeneity of ontology and data sources; entity linking is the basis of medical knowledge fusion, eliminating inconsistencies in knowledge through operations such as disambiguation; and relation deduction is used to discover implicit knowledge, thereby expanding and complementing the medical knowledge base.

2.4.1. Entity Alignment. Entity alignment eliminates inconsistencies such as entity conflict and unclear pointing in heterogeneous data, creating a large-scale unified knowledge base from the top layer, and helping machines understand multisource heterogeneous data and form high-quality knowledge. In medical big data, in the environment affected

Table 2: Commonly used data sets for medical entity relation extraction.

Name	Details
DrugBank	More than 80 aspects are provided for each drug, including brand name, chemical structure, protein and DNA sequences, related links on the Internet, feature descriptions and detailed pathological information, and so on.
STITCH	A platform for searching known and predicted interactions between compounds and proteins. The STITCH database contains more than 30,000 small molecular compounds and 2.6 million protein interactions from 1,133 species.
TCMSP	It includes 499 traditional Chinese medicine registered in Chinese Pharmacopoeia, including 29,384 ingredients, 3,311 targets, and 837 related diseases. This information can be queried and downloaded into the database. The disease information in this database comes from the TTD and PharmGKB databases.
TTD	Provide information about drugs, targets, diseases, and pathways. The current version collects 34,019 drugs, including 2,544 licensed drugs, 8,103 clinical trial drugs, and 18,923 drugs under development. Each drug provides information on its chemical structure, targets, targeted diseases, and related pathways. Users can search the database through targets, medications, conditions, and biomarkers and use drug similarity search tools to predict the targets of compounds without target information.
CCHMC	The data are from CCHMC (Cincinnati Children's Hospital Medical Center). CCHMC's institutional review committee approved the release of the data. All outpatient chest x-ray films and revisit chest films were sampled for one year by the bootstrap method. These data are commonly used data, and they are designed to provide sufficient code to cover the actual proportion of pediatric radiology activities.
MIMIC	A publicly available data set developed by the Computational Physiology Laboratory of the Massachusetts Institute of Technology, including unidentified health data (including demographics, vital signs, laboratory tests, medication, etc.) are related to about 40,000 intensive care patients.

by the size of the medical knowledge base, entity alignment will face the following three challenges:

(1) High computational complexity. The algorithm's computational complexity will increase quadratically with the size of the knowledge base, and the computational complexity is unacceptable.

(2) The quality of data varies. Due to the different construction purposes and methods of other medical knowledge bases, there may be similar repeated data, isolated data, and varying time intensity.

(3) Missing training data. Most medical databases do not have initial data, and researchers usually need to manually label the data to construct training data, which is also a huge task.

Existing entity alignment algorithms can be divided into pairwise entity alignment and collective entity alignment. Pairwise entity alignment methods only consider instances and their attribute similarity. Commonly used methods include probability and statistics model, regression classification tree model, support vector machine classification model, ensemble learning model, hierarchical graph model, and so on. The collective entity alignment method is based on the paired entity alignment and adds the interentity relationship when calculating the entity similarity. Standard techniques include vector space model, bootstrapping algorithm, Bayesian network model, LDA allocation model, Markov logic network model, and so on [35].

The most commonly used data sets for entity alignment are shown in Table 3.

Table 3: Commonly used data sets for entity alignment.

Name	Amount of data
Freebase FB15k	Entities: 14,951 Relationship: 1,345 Triple: 592,213
WK3l-15k	Entities: 60,293 Relationship: 7,087 Triple: 725,970
DBP15k	Entities: 498,765 Relationship: 12,874 Triple: 1,260,076
DWY100K	Entities: 400,000 Relationship: 883 Triple: 1,843,583

2.4.2. Entity Linking. The primary function of entity linking is to use the entities in the medical knowledge base to disambiguate the entity references obtained from the text of medical big data and identify each entity reference to its corresponding mapping entity in the medical knowledge base. The entity here refers to a textual representation of an entity [36]. A medical entity may have many different expressions, such as full name, alias, abbreviation, and so on. According to the information used by entity links, existing work is mainly divided into entity attributes-based entity linking method [37], entity popularity-based entity linking method [38], context-based entity linking method [39], and external evidence-based entity linking method [40]. The entity linking method classification table is shown in Table 4.

In constructing a knowledge base, entity recognition is the premise of entity linking, and entity recognition can provide more effective information. The joint learning of entity linking and entity recognition can reduce the workload. The joint solution of entity recognition and entity linking tasks can improve the performance of named entity recognition and entity linking, which is the focus and difficulty of current medical knowledge graph research.

In 2017, Lou et al. [41] proposed a disease entity recognition and normalization model, transforming the output construction process into a progressive state transition

Table 4: Classification table of entity linking methods.

Name	Rule	Advantages	Disadvantages
Entity attributes-based	Computing the similarity between the attributes	Simple realization and the high accuracy rate when the medical attribute is rich	With poor antinoise ability, the accuracy cannot be guaranteed when the entity attribute is sparse.
Entity popularity-based	Based on probability statistics and the frequency in the medical encyclopedia	Reliable and simple heuristic rules	Poor robustness and the ambiguity of specific entities are not considered.
Context-based	Based on the similarity between the entity contexts	High accuracy when the text is long enough and relatively clean	With less flexibility, the accuracy rate cannot be guaranteed when the text is sparse.
External evidence-based	Based on the semantic correlation	Introduced strong expansibility of rich feature information	Method effectiveness depends on the quality of external evidence.

process, allowing nonlocal features. Experiments show that the federated framework performs better than other methods of performing tasks separately. Compared with other advanced methods, this method has more advantages.

In 2019, Zhao et al. [42] proposed a new deep neural multitask learning framework with an explicit feedback strategy for joint entity recognition and entity normalization modeling. This method uses multitask learning to make a general representation of two tasks. It successfully converts cross-architecture tasks into parallel multitask settings while maintaining mutual support between tasks. The experimental results show that this method performed better than the most advanced method at that time on two open medical literature data sets.

2.4.3. Relationship Deduction. A preliminary ontology prototype can be obtained through entity alignment and entity linking. However, when building a knowledge base, different requirements and design concepts will lead to the diversity and heterogeneity of data in the knowledge base. Therefore, to form high-quality medical knowledge, it is also necessary to continuously carry out relationship deduction, expand the dynamically generated relationship into the existing medical knowledge base, form a large-scale medical knowledge system from the level, manage knowledge in a unified way, and improve the freshness and coverage of the medical knowledge base. It is very important. Due to the randomness of medical natural language expression, there are many synonymous or polysemous expressions in the relationship, which brings great challenges to expanding the relationship.

The main goal of relationship deduction is to dynamically extend the entity relations obtained from the medical big data text to the knowledge base. There are two possible situations of medical entity relations: (1) there is an entity relation in the medical knowledge base that is equivalent to the target text entity relation. You only need to find the entity relationship corresponding to the text entity relationship in the medical knowledge base. (2) If no entity relationship is the same as the target text entity relationship in the medical knowledge base, you need to extend the entity relationship into the knowledge base and complete the merging of the medical text entity relationship and the medical knowledge base entity relationship.

The key to relationship deduction is determining whether two entity relationships represent the same. There are currently two methods: (1) traditional semantic-based methods, which verify whether it is the same relationship by comparing the semantic similarity between words describing the relationship, and (2) methods based on embedding learning. This method is based on the embedding space. Find an appropriate energy function to learn the entity's embedded representation, use the entity's embedded representation to express the entity relationship, and judge whether the relationship describing the entity expresses the same relationship to realize the structural mapping of the entity relationship.

Although there have been some developments in the current knowledge fusion technology in the medical field, it still requires a lot of manual intervention, and efficient knowledge fusion algorithms still need to be studied.

2.5. Knowledge Reasoning. The reasoning is to mine hidden information from existing knowledge. In contrast, knowledge reasoning pays more attention to selecting and applying knowledge and methods, minimizing manual participation, deducing missing facts, and solving problems. In the medical knowledge graph, knowledge reasoning helps doctors collect patient data, diagnose and treat diseases, and control the rate of medical errors. However, doctors will make different diagnoses according to the patient's condition, even for the same disease. The medical knowledge graph must deal with repetitive and contradictory information, increasing the complexity of constructing the medical reasoning model.

Traditional knowledge reasoning methods include description logic reasoning [43], rule-based reasoning [44], case-based reasoning [45], and so on. Although traditional knowledge reasoning methods have promoted the development of medical knowledge graphs to a certain extent, there are also defects such as low accuracy, low data utilization, and insufficient learning ability, which do not meet the requirements of practical applications.

With the rapid growth of medical big data, there are some problems in traditional knowledge reasoning methods, such as missing information, prolonging diagnosis time, and so on. Artificial intelligence technology has natural advantages for useful mining information from massive medical

data and can improve the efficiency and accuracy of knowledge reasoning. Standard models include artificial neural networks model [46], genetic algorithm [47], and backpropagation. The development and application of artificial intelligence technology have improved the construction efficiency of medical knowledge graphs and the accuracy of knowledge reasoning. Medical knowledge graph must deal with many repeated and contradictory medical information. For example, doctors should make different diagnoses according to the patient's condition [48] and give different solutions. Artificial intelligence has the natural advantage of useful mining information from massive data. Knowledge reasoning involves selecting and applying knowledge and methods and can infer the problem's solution [49]. Traditional and artificial intelligence reasoning methods take the knowledge graph as the data source.

In contrast, graph-based reasoning regards the knowledge graph as a graph, the medical entity as a node, and the relationship between entities as an edge. Using the information contained in the relationship path, the semantic relationship between the two entities is analyzed through the multistep path between the two entities. Standard algorithms include path-constraint random walk algorithm, path ranking algorithm, and so on.

2.6. Construction of Medical Knowledge Map of Chronic Diseases. This article uses a chronic disease knowledge graph as an example to introduce the construction of medical knowledge graph. Medical knowledge graph construction methods are generally divided into three types: top-down, bottom-up, and a combination of the above two methods. The top-down approach builds the top-level relational ontology and then updates the extracted entity matching to the constructed top-level ontology. The bottom-up approach directly incorporates the categories, entities, attributes, and relationships in the extracted data into the knowledge graph. No matter which method is used to build the knowledge graph, its construction flow chart is shown in Figure 1.

The construction of a chronic disease knowledge map requires integrating multiple resources, such as medical research literature on chronic diseases, chronic disease health care knowledge, and clinical diagnosis and treatment case database data. These data are distributed in many places such as the Internet, scientific literature databases, specialist diagnosis and treatment datasets, paper books, and so on, and these resources are of various types, with different focuses and qualities. To build a knowledge map about chronic diseases, it is necessary to consider the comprehensiveness and high quality of the system. Therefore, filtering and extracting numerous resources is necessary to make them structured and systematic. At the same time, to realize the effective organization and expansion of chronic disease knowledge, it is necessary to standardize the medical terminology involved in chronic disease knowledge to ensure that chronic disease-related data can be effectively collected and processed.

Most of the treatment knowledge of chronic diseases comes from the clinical experience of doctors and the established effective methods for treating chronic STDs. This difference in medical experience and knowledge in specific medical diagnoses and treatments brings difficulties to standardizing chronic disease knowledge. Currently, there is no standardized system of chronic disease pathology, incidence law, medical control, or other resources. Therefore, to realize the systematization of knowledge related to chronic disease treatment, it is necessary to collect, classify, and organize chronic disease knowledge resources, irrelevant filter content, and build a knowledge map about chronic disease diagnosis and treatment to realize the intelligent medical treatment of chronic diseases.

The process of establishing chronic disease knowledge extraction is shown in Figure 2. First, comprehensively collect various resources about chronic diseases and use digital technology to digitize them. For complex professional medical experience knowledge, the forms include various pictures, audio, and video resources, and they are classified and summarized by chronic disease experts to make them structured. Systematized and systematized to form a collection of knowledge about the diagnosis and treatment of chronic diseases. Under the relevant medical terminology standards, perform text analysis on textual chronic disease resources, such as semantic labeling, entity, and relationship extraction, association analysis, and so on. Differences in content and types of relationships between concepts constitute a top-level ontology about chronic diseases. Finally, the chronic disease knowledge collection and top-level ontology are combined to build a knowledge base about chronic diseases.

The construction of a chronic disease knowledge map includes two parts: chronic disease treatment experience knowledge base and chronic disease domain ontology. The chronic disease treatment experience knowledge base is the knowledge about various causes, effective drugs, treatment processes, treatment cycle, contraindications, and so on of chronic disease treatment, which is the core of the chronic disease knowledge base; the chronic disease domain ontology is the medical concept and concept of chronic disease The term set of interrelationships is designed to meet the needs of sharing and expanding the chronic disease knowledge base, most of which can be achieved by referencing existing clinical term sets or disease ontology.

2.7. Ethical Approval. Ethical approval is not necessary because no human subjects and patient information were collected and studied.

3. Discussion

Knowledge graph technology is a transformation and sublimation of semantic web technology. Since Google put forward the knowledge graph concept, its popularity has only increased. Through the in-depth observation and analysis of the technical system of knowledge graph construction, it can be seen that it is a practical technology based

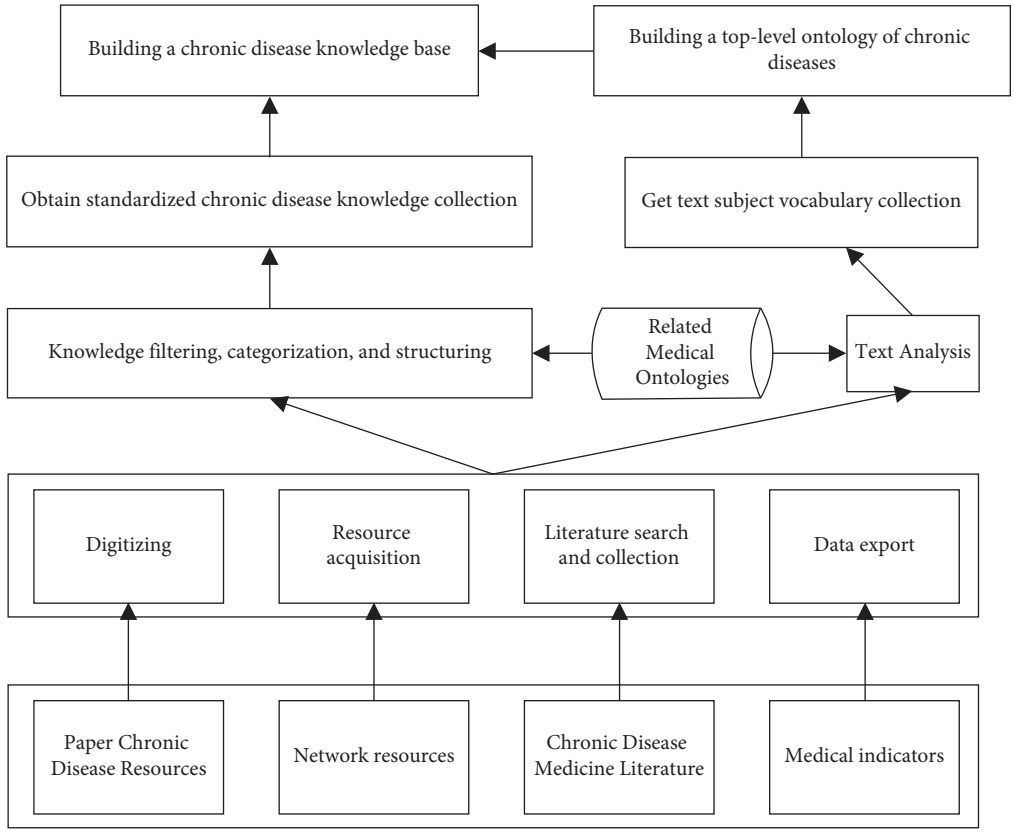

Figure 2: Chronic disease knowledge extraction and establishment process.

on the research achievements of multidisciplinary fields. It is a collection of theoretical research hotspots and application technologies in artificial intelligence, information retrieval, natural language processing, the World Wide Web, and so on. As far as the medical field is concerned, because of the professionalism, standardization, and limited terminology of medical knowledge and rules, high-quality data can be obtained from standardized medical dictionaries, and medical databases and medical knowledge maps can be constructed. The general development direction of the medical knowledge graph in the future should reflect the following aspects:

3.1. Multilingual Medical Knowledge Graph. The mutual integration of medical knowledge at home and abroad is more conducive to the development of the medical field, and multilingual medical knowledge graph technology is the key to realizing the mutual communication and exchange of medical knowledge within different national boundaries. This will become an important trend in developing medical knowledge graphs in the future.

3.2. Large-Scale Multimodal and Multisource Medical Knowledge Base. Affected by many factors, the scale of the existing medical knowledge graph is mainly limited, and the mode of expression is relatively simple, mainly in text and graph data. Still, sound, images, pictures, and so on also contain much medical information. There are also many medical images, x-rays, and other multimodal information in a medical clinic. The sources of medical knowledge can also come from books, literature, web pages, videos, and so on. Therefore, a hot spot in future medical knowledge graph research is to build a large-scale multimodal and multisource medical knowledge base.

3.3. Cognitive Knowledge Graph. The cognitive knowledge graph studies the models and methods of large-scale multigranularity knowledge reasoning based on the mutual constraints of deep learning and logical reasoning and develops a large-scale knowledge reasoning system based on the combination of ontology, rules, and deep understanding, which enables it to reason the knowledge base and constraints containing billion-level RDF (resource description framework) triple. The average response time is in seconds and has good scalability. The knowledge evolution model and prediction method based on spatiotemporal characteristics are studied, and a knowledge evolution system is developed to update the knowledge base in real-time. The average response time is in seconds.

3.4. Visual Knowledge Graph. The real significance of visualizing a knowledge graph is to make people understand the process and result of reasoning visually. On the other hand, the visualization of medical knowledge graph is from the standpoint of doctors or patients to seek the best knowledge display scheme: patients can understand the

diagnosis results, and doctors can make a reasonable diagnosis using the dynamic reasoning process of the knowledge graph.

3.5. Challenges Faced by Medical Knowledge Graph

(1) Text extraction is difficult. In medical knowledge extraction, the research on knowledge extraction methods for the open domain is still in its infancy. Although some research results have achieved good results on specific data sets, there are problems such as low algorithm accuracy, many constraints, and poor scalability. In particular, the extraction of pure text information involved in the extraction of medical electronic medical records is an important challenge currently facing.

(2) Knowledge graph storage method. At present, the medical knowledge graph is mainly stored in a graph database. While benefiting from the query efficiency brought by the graph database, it will also lose the advantages of the relational database, such as the graph database cannot support SQL language query, and the query efficiency is low. Translating natural language query sentences into query expressions and equivalent expressions that knowledge graphs can understand is also a key problem to be solved in applying medical knowledge graphs.

(3) The entity correspondence is inaccurate. The main challenge in the medical knowledge fusion stage is to achieve accurate entity linking. Although the research on entity disambiguation and coreference resolution technology has a long history, medical entities have serious multisource referential problems in different data sources due to the diversity of medical knowledge sources. The research results obtained are far from practical application in the medical field. The current academic circles are a common concern in correctly connecting the entities extracted from the text to the medical knowledge base under cross-context and cross-text conditions.

4. Conclusions

Electronic medical data has accumulated to a certain extent with the development of medical information. The construction of the knowledge graph in the medical field can extract medical knowledge from massive data and manage, share, and apply it reasonably and efficiently, which is of great significance to today's medical industry. From the perspective of the construction of medical knowledge graph, this paper makes a comprehensive investigation and in-depth analysis of the structure, key technologies, research, and application status of medical knowledge graph construction. It looks forward to its future research direction. The significance of the knowledge graph in the medical field is that it is a global medical knowledge base and the basis for supporting intelligent medical applications such as auxiliary diagnosis and treatment, intelligent search, and so on. Medical knowledge graph combines knowledge graph with medical knowledge, promoting the automatic and intelligent processing of medical data and bringing new development opportunities for the medical industry. Although there are many meaningful attempts in the research of medical knowledge graphs, it is not perfect and in-depth, and further research is needed.

Abbreviations

RDF:	Resource description framework
WOL:	Web ontology language
RDFS:	Resource description framework schema
DAML:	DARPA Agent Markup Language
SNOMED-CT:	Systematized Nomenclature of Medicine—Clinical Terms
ICD:	International classification of diseases
UMLS:	Unified medical language system
SNERL:	Simultaneous neural entity-relation linker
APGC:	AI professional gemini cases
IBM:	International Business Machines Corporation.

Data Availability

No data were used to support this study.

Disclosure

The paper does not involve human or animal research.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[2] J. Mervis, "Agencies rally to tackle big data," *Science*, vol. 336, no. 6077, p. 22, 2012.

[3] G. B. Orgaz, J. Jung, and D. Camacho, "Social big data: recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[4] V. Mayerschnberger and K. Cukier, "Bigdata: arevolution thatwil transform how we live, work, and think [J]," *Mathematics and Computer Education*, vol. 47, no. 17, pp. 181–183, 2014.

[5] K. Yuan, Y. Deng, D. Chen, B. Zhang, K. Lei, and Y. Shen, "Construction techniques and research development of medical knowledge graph," *Application Research of Computers*, vol. 8, no. 7, pp. 1929–1936, 2018.

[6] R. Tong, C. Sun, and H. Wang, "Construction of traditional Chinese, medical, knowledge, graph and its application," *Journal of Medical Information*, vol. 37, no. 4, pp. 8–13, 2016.

[7] T. B. Murdoch and A. S. Detsky, "The inevitable Application of big data to health care," *The Journal of the American Medical Association*, vol. 309, no. 13, p. 1351, 2013.

[8] E. H. Shortliffe, *Computer-based Medical Consultations: MYCIN*, Elsevier Publishing Co. Inc, New York, NY, U.S.A, 2012.

[9] G. P. Rédei, *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, Springer, Netherlands, 2008.

[10] W. Ceusters, P. Martens, C. Dhaen, and B. Terzic, "LinkFactory: an advanced formal ontology management system," in *Proceedings of the of Interactive Tools for Knowledge Capture Workshop*, pp. 175–204, January 2001, https://www.researchgate.net/publication/244468211_LinkFactory_an_Advanced_Formal_Ontology_Management_System.

[11] P. G. Baker, A. Brass, S. Bechhofer, C. A. Goble, N. W. Paton, and R. Stevens, "TAMBIS: transparent access to multiple bioinformatics information sources," in *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, pp. 25–34, AAAI Press, Palo Alto CA, U.S.A, June 1998.

[12] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning Structured Embeddings of Knowledge bases," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 301–306, AAAI Press, San Francisco, CA, U.S.A, January 2011.

[13] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 926–934, Curran Associates Inc, Red Hook, NY, U.S.A, January 2013.

[14] R. Jonathon, N. L. Roux, A. Bordes, and G. R. Obozinski, "A Latent Factor Model for Highly Multi-Relational data," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 3167–3175, Curran Associates Inc, Red Hook, NY, U.S.A, December 2012.

[15] A. Bordes, N. Usunier, A. G. Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-Relational data," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 2787–2795, Curran Associates Inc, Red Hook, NY, U.S.A, June 2013.

[16] D. Kleyko, S. Khan, E. Osipov, and S. P. Yong, "Modality Classification of Medical Images with Distributed Representations Based on Cellular Automata Reservoir computing," in *Proceedings of the 14th IEEE International Symposium on Biomedical Imaging*, IEEE Press, Piscataway, NJ, U.S.A, April 2017.

[17] A. Henriksson, J. Zhao, H. Dalianis, and H. Boström, "Ensembles of randomized trees using diverse distributed representations of clinical events," *BMC Medical Informatics and Decision Making*, vol. 16, no. 2, p. 69, 2016.

[18] F. McDonald and P. L. Elkin, "UMLS concept indexing for production databases: a feasibility study," *Journal of the American Medical Informatics Association*, vol. 8, no. 5, pp. 512–514, 2001.

[19] D. Han, H. LlQ, and W. CAl, "Research and application of artificial intelligence in medical imaging," *Big Data Research*, vol. 5, no. 1, pp. 39–67, 2019.

[20] S. T. Wu, H. Liu, D. Li et al., "Unified medical language system term occurrences in clinical notes: alarge-scalecorpusanalysis," *Journal of the American Medical Information Association Jamia*, vol. 19, no. 1, pp. 149–156, 2012.

[21] C A. Smith and P. Z. Stavri, *Consumer Health vocabulary*, pp. 122–128, Springer Publishing, New York, NY, U.S.A, 2005.

[22] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[23] R. I. Dogan, R. Leaman, and L. Zhiyong, "NCBI disease corpus: a resource for disease name recognition and concept normalization," *Journal of Biomedical Informatics*, vol. 47, no. 2, pp. 1–10, 2014.

[24] J. Kazama, T. Makino, Y. Ohta, and J. I. Tsujii, "Tuning support vector machines for biomedical named entity recognition," in *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Stroudsburg: Association for Computational Linguistics*, pp. 1–8, Philadelphia, PA, USA, July 2002.

[25] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," *Bioinformatics*, vol. 20, no. 7, pp. 1178–1190, 2004.

[26] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[27] A. N. Jagannatha and Y. Hong, "Structured Prediction Models for Labeling in Clinical Research," in *Proceedings of the 2016 Conf on Empirical Methods in Natural Language Processing*, pp. 856–865, ACL, New York, NY, U.S.A, November 2016.

[28] C. Zhang, L. Yaliang, D. Nanx, W. Fan, and P. S. Yu, "On the generative discovery of structured medical knowledge," in *Proceedings of the ACM SIGKDD Int Conf on Knowledge Discovery& Data Mining*, pp. 23–37, ACM, New York, NY, U.S.A, July 2018.

[29] A B. Abacha and P. Zweigenbaum, "MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies," *Information Processing & Management*, vol. 51, no. 5, pp. 570–594, 2015.

[30] M. Eberts and A. Ulges, "Span-based Joint Entity and Relation Extractionwith Transformer pre-training," 2019, https://arxiv.org/abs/1909.07755.

[31] T. Bansal, P. Verga, N. Choudhary, and M. Andrew, "Simultaneously Linking Entities and Extracting Relations from Biomedical Text without Mention-Level supervision," 2019, https://arxiv.org/abs/1912.01070.

[32] G. s Nan, Z. J. GUo, I. SekuliC, and W. Lu, "Re Asoning with Latent Structurerefinement for Document-Level relationextraction," 2020, https://arxiv.org/abs/2005.06312.

[33] H. Lin, Y. Wang, Y. Jia, P. Zhang, and W. P. Wang, "Network bigdata oriented knowledge fusion methods: a survey," *Chinese Journal of Computers*, vol. 23, no. 1, pp. 1–27, 2017.

[34] X. L. Dong, W. Horn, K. Murphy et al., "From data fusion to knowledge fusion," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 881–892, 2014.

[35] S. z Kang, L. X. Ji, Z. Li, X. Hao, and Y. Ding, "Iterative cross-lingual entity alignment based on TransC," *IEICE - Transactions on Info and Systems*, vol. 103, no. 5, pp. 1002–1005, 2020.

[36] M. Bilenko and R. J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," in *Proceedings of the ACM IntConf on Knowledge Discovery and Data Mining*, pp. 39–48, ACM, New York, NY, U.S.A, August 2003.

[37] R. C. Chen, C. Bau, and C. J. Yeh, "Merging domain ontologies based on the WordNet system and fuzzy formal concept analysis techniques," *Applied Soft Computing*, vol. 11, no. 2, pp. 1908–1923, 2011.

[38] Y. Li, C. Wang, H. Fangqiu, J. Han, D. Roth, and X. Yan, "Mining Evidence for Named Entity disambiguation," in *Proceedings of the ACM SIGKDDInt Conf on Knowledge Discovery and Data Mining*, pp. 1070–1078, ACM, New York, NY, U.S.A, August 2013.

[39] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM transactions on knowledge discovery from Data*, vol. 1, no. 1, pp. 299–304, 2007.

[40] Y. Shen, Y. Deng, M. Yang et al., "Knowledge-aware attentive neural network for ranking question answer pairs," in

Proceedings of the 41st Int ACM SIGIR Conf on Research 8. Development in Information Retrieval, pp. 901–904, ACM, New York, NY, U.S.A, June 2018.

[41] Y. X. Lou, F. Li, S. Xiong et al., "A transition-based joint model for disease named entity recognition and normalization," *Bioinformatics*, vol. 33, no. 15, pp. 2363–2371, 2017.

[42] S. D. Zhao, T. Liu, F. Wang, and S. C. Zhao, "A neural multi-task learning framework to jointly model medical named entity recognition and normalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 817–824, Honolulu, U.S.A, January 2019.

[43] A. Mourao, F. Martins, and J. Magalhaes, "Multimodal medical information retrieval with unsupervised rank fusion," *Computerized Medical Imaging and Graphics*, vol. 39, pp. 35–45, 2015.

[44] B. Buchanan and E. H. Shortliffe, *Rule-based expert systems: the MYCIN experiments of the stanford HeuristicProgramming project*Vol. 67, Addison-Wesley, Boston, MA, U.S.A, 2013.

[45] C. Bousquet, C. Henegar, A. L Louet, P. Degoulet, and M. C. Jaulent, "Implementation of automated signal generation in pharmacovigilance using acknowledge-based approach," *International Journal of Medical Informatics*, vol. 74, no. 7, pp. 563–571, 2015.

[46] J. Khan, J. S. Wei, M. Ringner et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[47] D. E. Goldberg, *Genetic algorithm in search optimization and machine learning*, vol. 8, pp. 2104–2116, , no. 7, Addison-Wesley, Boston, MA, U.S.A, 1989.

[48] E. Ccwa, B. Wcy, C. Wdh et al., "Prediction of fatty liver disease using machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23–29, 2019.

[49] A. C. Lyngdoh, N. A. Choudhury, and S. Moulik, "Diabetes Disease Prediction Using Machine Learning Algorithms," in *Proceedings of the 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) - IECBES 2020*, March 2020.