STM JOURNALS

Research & Reviews: A Journal of
Medical Science and Technology

https://journals.stmjournals.com/rrjomst/

ISSN: 2319-3417 (Online)
ISSN: 2349-1272 (Print)
Volume 15, Issue 1, 2026
January–April
DOI (Journal): 10.37591/RRJOMST

Research RRJOMST

A Knowledge Graph Approach for Breast Cancer Diagnosis and Data Sharing Platform Implementation in the Context of Human Papillomavirus Infection

Xiang Fu[1], Lin Hu[2], Xuebei Du[3,*]

Abstract

Background: *Breast cancer remains among the most prevalent malignancies in women worldwide, and effective diagnosis and data integration continue to challenge clinical practice. Diagnostic reports from mammography and ultrasound contain rich clinical information that is often under-utilized due to heterogeneous formats and limited data-sharing infrastructure. In the context of human papillomavirus (HPV) infection, which may influence oncogenic pathways and data complexity, advanced computational methods offer new solutions to this problem.* ***Objective:*** *We aimed to construct a comprehensive knowledge graph for breast cancer diagnosis, derived from imaging reports, and to implement a data-sharing platform to enable structured, interoperable data management in a clinical oncology setting.* ***Methods:*** *We collected 1,002 mammography and ultrasound reports and employed a joint entity- and relationship-extraction pipeline using a RoBERTa-Global Pointer model. Model performance metrics accuracy, recall and F1-score exceeded 95%. Extracted information was loaded into a Neo4j knowledge graph containing 23,421 entities and 12,589 relationships. A Spring Boot-based web platform provided interactive visualization, advanced search and intelligent data management of breast cancer and HPV-related diagnostic information.* ***Results:*** *The system demonstrated excellent extraction performance and established a rich semantic network that supports structured data exchange, query and analysis. The platform supports clinicians and researchers in linking imaging features, HPV-related metadata and therapeutic decision-making in a unified data environment.* ***Conclusions:*** *Our knowledge graph and data-sharing platform provide a robust and scalable framework for breast cancer diagnosis and management addressing data-format heterogeneity, enhancing interoperability and enabling precision-driven clinical workflows. This approach offers a generalizable template for other disease domains, particularly where complex imaging and infection-related data converge.*

***Author for Correspondence**
Xuebei Du
E-mail: xuebeiD@outlook.com

[1]Researcher, Department of Information Center, Longgang Central Hospital of Shenzhen, Shenzhen City, 518116, Guangdong Province, China
[2]Researcher, School of Electronic Information and Electrical Engineering, Chengdu University, Chengdu, 610106, Sichuan Province, China
[3]Researcher Department of Geriatric Infection, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, 610072, Sichuan Province, China

Received Date: Feb 03, 2026
Accepted Date: Feb 13, 2026
Published Date: April 29, 2026

Citation: Xiang Fu, Lin Hu, Xuebei Du. A Knowledge Graph Approach for Breast Cancer Diagnosis and Data Sharing Platform Implementation in the Context of Human Papillomavirus Infection. Research & Reviews: A Journal of Medical Science and Technology. 2026; 15(1): 75–97p.

Keywords: Breast cancer, knowledge graph, human papillomavirus, breast mammography, data sharing platform implementation.

INTRODUCTION

Breast cancer is one of the most common malignant tumors threatening women's health. The five-year prevalence statistics indicate that approximately 11% of breast cancer cases occur in China, and over the past few decades, the incidence rate has risen rapidly. Therefore, convenient early diagnosis and timely treatment are crucial for improving patients' prognosis and prolonging their survival period, especially in regions with poor medical resources [1, 2]. Breast ultrasound and mammography, which are fast, noninvasive, low-

© STM Journals 2026. All Rights Reserved

cost, widely available, and suitable for individuals across different income levels and regions, are essential methods for the early screening and diagnosis of breast diseases [3]. Mammography is particularly sensitive to calcifications in breast tissue but is less sensitive to detect dense breast cancers without calcifications, whereas ultrasound examination has high sensitivity for hypoechoic areas within dense soft tissues but is less sensitive to calcified tissues, which might lead to missed diagnoses [4, 5]. Therefore, a growing body of research has revealed that combining breast ultrasound with mammography can effectively reduce missed diagnoses of breast cancer [6].

With the continuous development of artificial intelligence technology, its application value in the medical field has become gradually prominent, providing more convenient means for medical information retrieval and sharing. Among these technologies, the knowledge graph has gradually attracted attention as a semantic network that visually illustrates relationships between entities through nodes representing the entities and edges demonstrating their connections [7]. In the medical field, knowledge graphs have tremendous potential. They not only filter, integrate, and present medical information, building databases of medical knowledge relationships but also develop systems for auxiliary diagnosis, risk prediction, and clinical decision support [8]. Most general medical knowledge graphs are constructed based on a single data source, neglecting to effectively integrate high-quality medical knowledge from various sources. Furthermore, most of them focus on general medical knowledge (An, 2023), but seldom address specific diseases with real clinical data [9].

However, the existing breast cancer-related knowledge graphs still have drawbacks. Although several breast cancer knowledge graphs have been constructed by integrating knowledge from clinical guidelines, medical encyclopedias, and electronic medical records, they still lack in-depth analysis and vary significantly in quality due to the diversity and complexity of these data. Another research utilized semantic and evidence-based biomedical knowledge to set up the risk-explorer knowledge graph, which demonstrated the causal connection between risk factors and breast cancer. However, current research on breast cancer knowledge graphs is limited and lacks effective resource integration. Therefore, there is an urgent need to design an effective date processing and management system platform to enable collection, precise management, and in-depth analysis of breast cancer variables [10].

In constructing a knowledge graph, acquiring the data layer is crucial, with the extraction of entities and relationships being an essential component. Compared to general fields, the extraction of entities and relationships in the biomedical field is more complex. Traditional models are not applicable to Chinese language environment. Existing models have limitations in dealing with remote semantic information, nested structures and data sources [11]. Pre-trained BERT on Chinese clinical texts and integrated its results with the BiLSTM-CRF model, specifically addressing the issue of named entity recognition for breast cancer, achieving an F1 score of 93.53%. Liu et al. further developed the RoBERTa model, making it better than the BERT model in terms of training data volume, batch size, and model parameters, which in turn improved the BERT model's performance. It is worth noting that the RoBERTa model shows greater effectiveness than the traditional models and their improved versions in the joint extraction of entities and relationships. Considering the above issues, our approach used combined breast cancer ultrasound and mammography examination reports as research samples. We extracted diagnostic knowledge related to breast cancer from multi-source databases and developed a joint extraction model of entities and relationships using RoBERTa and Global Pointer. The model was then used to construct a breast cancer diagnostic knowledge graph, with the primary function being the development of a data-sharing service platform. Leveraging the breast cancer diagnostic knowledge graph and electronic medical records, we developed a user-friendly, intuitive cloud service platform to achieve effective resource integration and utilization. This platform enables doctors and researchers to access and analyze medical record information more effectively, setting new directions for future breast cancer research [12]. The main contributions of this study are as follows:

Unlike previous studies that focused solely on either ultrasound or mammography reports, this study constructs a knowledge graph that integrates features from both, aiming to capture a broader spectrum of breast cancer diagnostic knowledge [13].

© STM Journals 2026. All Rights Reserved

This study proposes the construction of a breast cancer diagnostic knowledge graph using a joint extraction model of entities and relationships based on RoBERTa and Global Pointer. Compared to traditional sequential extraction models, this approach offers greater accuracy and efficiency. (A combined top–down and bottom–up approach was used to construct the ontology. Compared to traditional methods that use either top–down or bottom–up construction, this combined approach better facilitates the discovery, standardization, and procedural development of new knowledge) [14].

In terms of applications, the knowledge graph-based cloud service platform can provide functionalities like visual analysis, auxiliary diagnosis, intelligent retrieval, personalized recommendations, and intelligent question-answering. These features enhance its practicality and effectiveness in clinical applications.

METHODS

Data Acquisition and Curation

This research was approved by the hospital ethics committee of Longgang Central Hospital approval number LCH 0024001. A total of 1,084 reports, including those of 504 female breast cancer patients who underwent both ultrasound and mammography examinations, were collected from October 2019 to August 2021 at the medical information center of a tertiary hospital in China. We removed reports with incomplete or incorrect data, performed data deduplication, handled missing values, converted between simplified and traditional Chinese characters, cleaned invalid information, and anonymized the data. Ultimately, 1,002 reports were extracted for constructing the breast cancer diagnostic knowledge graph [15–20].

Knowledge Graph Construction Framework

The overall process for constructing the knowledge graph primarily includes the following four steps and is shown in Figure 1:

- *Step One:* Design of breast cancer diagnostic knowledge graph ontology and conceptual layer.
- *Step Two:* Acquisition of the data layer.
- *Step Three:* Construction of the knowledge graph.
- *Step Four:* The application of the knowledge graph: data sharing platform implementation.

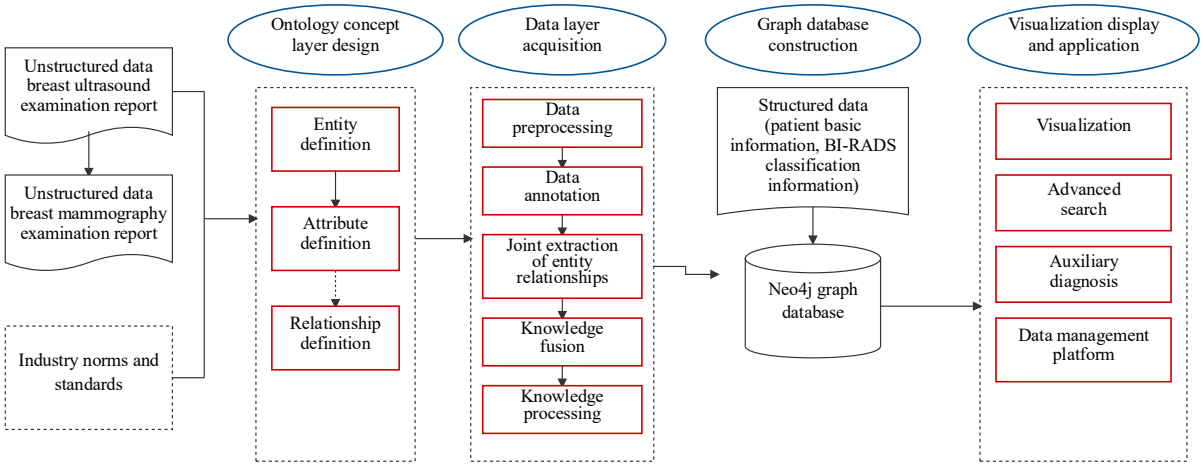


Figure 1. Overall process of knowledge graph construction.

Conceptual Layer and Ontology Design

At the logical architecture level, the knowledge graph is divided into two layers: the conceptual layer (or schema layer) and the data layer (or instance layer), as shown in Figure 2. As a crucial component of the knowledge graph, the conceptual layer forms the foundation for the data layer by defining the hierarchy and structural relationships among knowledge elements, including entities, relationships, and attributes. Entities represent the fundamental units of knowledge, relationships describe the connections between these entities, and attributes provide specific details about the entities [21–24]. We relied on the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS)

© STM Journals 2026. All Rights Reserved

(Radiology, 2003) and the Guidelines and Standards for the Diagnosis and Treatment of Breast Cancer by the Chinese Anti-Cancer Association (2021 Edition). Additionally, we defined the conceptual layer based on suggestions from professional physicians and multiple hospital reports and then revised it according to extensive domestic and international literature and industry standards. Ontology construction utilizes a combination of both top–down and bottom–up methods. During this phase, we referenced breast cancer clinical treatment guidelines and compiled 50 pairs of entity relationships, which were then subjected to expert evaluation [25–29].

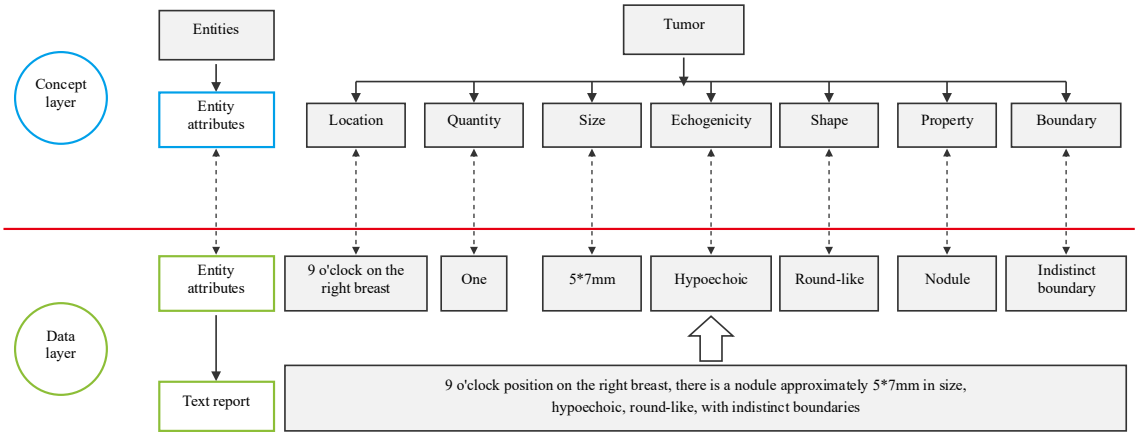


Figure 2. Corresponding relation of the conceptual layer and data layer.

The conceptual layer is designed as a three-level hierarchical structure. The first level contains 7 entities such as "common signs", the second level includes 17 entities such as "tumor", and the third level consists of related entity attributes (Table 1) [30].

Table 1. Entities and attributes of the conceptual layer in the breast cancer diagnostic knowledge graph.

Project	Primary entity	Secondary entity	Entity attributes
Ultrasound Examination Features	Common Signs	Mass	Shape, Size, Edge Features, Internal Echo, Echo Uniformity, Posterior Echo, Description, Location, Type.
		Calcifications	Quantity, Shape, Distribution, Type, Location, Negation.
		Surrounding Tissue	Description.
	Special Signs	Lymph Nodes	Location, Description, Negation, Echo.
	Associated Signs	Blood Flow	Location, Negation, Description, Blood Flow Speed.
		Skin	Shape, Location, Thickness.
		Duct Changes	Size, Description, Location.
		Structure	Distribution.
Mammography Examination Features	Common Signs	Mass	Shape, Size, Density, Features, Number, Location, Negation.
		Calcifications	Location, Number, Shape, Type, Size, Density, Negation, Distribution.
		Gland	Classification.
	Special Signs	Duct Changes	Distribution.
		Lymph Nodes	Location, Distribution, Number, Shape, Negation, Description.
		Heterogeneity	Location, Description, Negation, Echo.
	Concomitant Signs	Vascular Changes	Location, Distribution, Negation.
		Axillary Lymph Nodes	Number, Shape, Negation, Location, Description.
		Combined Abnormalities	Number, Location, Negation, Description.
	Category Description		Description, Location.

© STM Journals 2026. All Rights Reserved

In accordance with the principles of knowledge graph construction, it is necessary to delineate the contents of entities, relationships and attributes. The information about the subject, predicate, and object elements in the conceptual layer, illustrating the hierarchical relationship between entities and attributes. With the definition of these three knowledge elements, the construction of the conceptual layer of the breast cancer diagnosis knowledge graph was completed (Table 2).

Table 2. Three elements of the conceptual layer in the knowledge graph.

Subject	Predicate	Object
Entity (Basic Information)	select	Attribute (Name, Gender, etc.).
Entity (Report)	Has a	Entity (Basic Information, Breast Information, BI-RADS Information).
Entity (Breast Information)	Instance of	Primary Entity (Common Signs, Concomitant Signs, etc.).
Primary Entity (Common Signs, Concomitant Signs, etc.)	Part of	Secondary Entity (Mass, Calcifications, Structure, etc.).
Secondary Entity (Structure, Mass, etc.)	select	Attribute (Location, Size, etc.).
Feature Comparison	analyzes	Differences and similarities in the features of masses, lymph nodes, duct changes, and calcifications.
Overall Assessment	evaluates	Diagnostic judgment based on the combined results of ultrasound and mammography.

Data Layer Annotation and Entity-Relation Extraction

We established annotation standards and enhanced the consistency of annotation results through pre-annotation. Data annotation is divided into three stages: standard setting, pre-annotation, and formal annotation. Under the guidance from professional doctors, annotation standards were formulated to direct the subsequent annotation work. This study utilized a web-based annotation tool, ensuring the quality of data annotation through inter-annotator agreement (IAA) evaluation and expert review (an example of the annotation) (Figure 3). After data annotation, we built mapping relationships, performed data alignment, and finally transformed label sequences (using the BIO format: B indicates the beginning tag of an entity, I indicate inside an entity, and O indicates outside an entity) [31, 32].

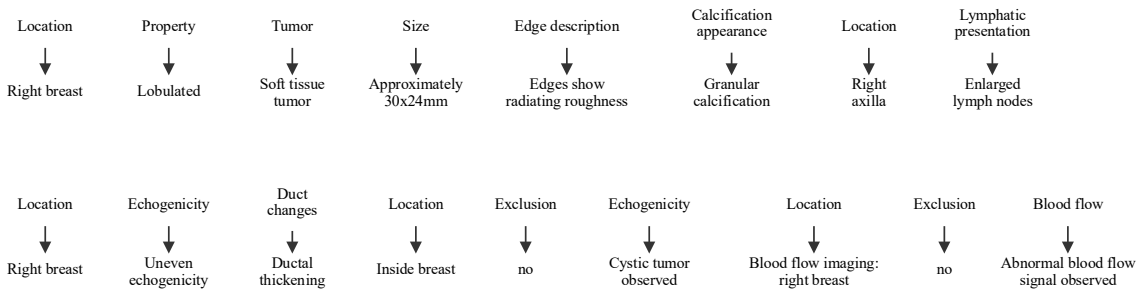


Figure 3. The example of data annotation.

Joint Entity and Relationship Extraction Model

We established annotation standards and enhanced the consistency of annotation results through pre-annotation. Data annotation was carried out according to these standards, followed by joint extraction of entities and relationships. The joint extraction of entities and relationships was the core component in the data layer, involving named entity recognition and relation extraction. Unlike traditional sequential extraction models, we developed a RoBERTa pre-trained model combined with Global Pointer to accomplish the joint extraction of entities and relationships [33].

As a modified deep learning model of BERT, RoBERTa model is specially designed for natural language processing tasks. Proposed by the Facebook AI research team in 2019, RoBERTa inherits the basic architecture of BERT, which is based on the Transformer architecture. RoBERTa meaningfully

© STM Journals 2026. All Rights Reserved

enhances the model's comprehension and generalization capabilities by optimizing the details of the pre-training process. Global Pointer is a decoding method based on span classification. It constructs an upper triangular matrix, with each element representing a possible span, and assigns an entity label to each span. By utilizing rotary position encoding, Global Pointer can account for the relative positions of entities within the report. This optimizes entity boundary predictions, leading to more accurate entity identification and resolving the issue of overlapping entities. Leveraging RoBERTa's strong language understanding capabilities, this model captures deep semantic information in the text. Additionally, the Global Pointer decoding layer effectively handles the crossover and nested relationships between entities, accurately identifying entities and their relationships within the text. The combination of RoBERTa and Global Pointer leverages the strengths of both methods, significantly enhancing the accuracy and efficiency of entity and relationship extraction [34, 35].

The architecture of the joint extraction model is shown in Figure 4. Specifically, the process flow is divided into the following steps:

- *Input Layer:* The model's input consisted of textual data from breast cancer ultrasound and mammography examination reports. This input encompassed patient basic information, along with detailed descriptions of lesion locations and characteristics.
- *Transformer Layer (RoBERTa):* The input text was first processed through the Transformer layers of the RoBERTa model, specifically a pre-trained Chinese version consisting of 12 Transformer layers. These layers encoded the text into high-dimensional feature representations, capturing lexical, grammatical, and contextual information. After processing by the RoBERTa model, each word or token in the text was represented by a corresponding vector. These representations, rich in semantic information, provided a foundation for the subsequent recognition of entities and relationships.
- *Global Pointer Decoding Layer:* This output was subsequently decoded using Global Pointer, a mechanism adept at complex entity recognition tasks such as resolving overlapping entities. By predicting the existence and type of relationships between any two text positions, Global Pointer performs joint extraction, thereby ensuring the accurate and efficient identification of entities and their interconnections.
- *Output Layer:* The model's output included the identified entities and their relationships. In this context, the output might comprise entities, such as basic patient information, lesion characteristics, and lesion locations, as well as the relationships between these entities such as the association between a lesion and its specific location.

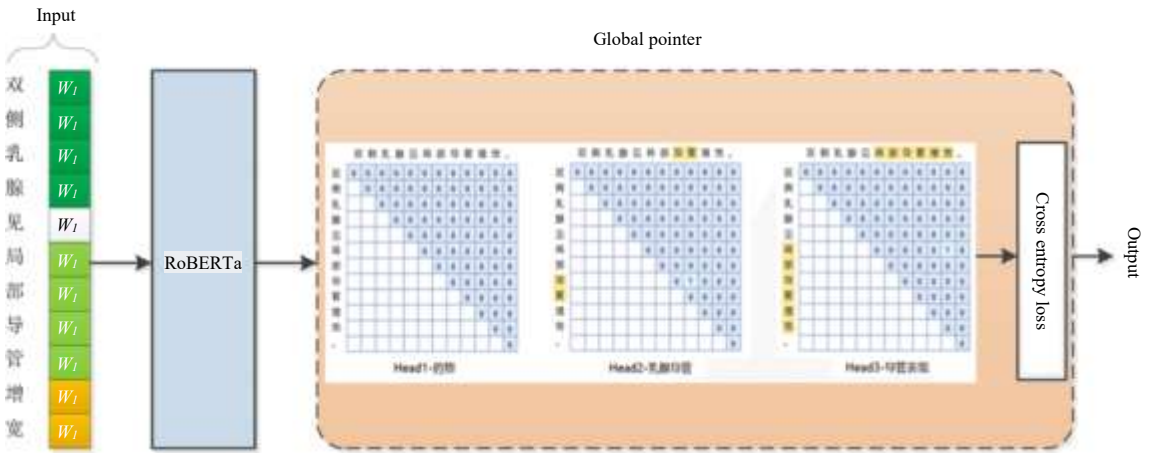


Figure 4. The framework for the joint extraction of entities and relationships.

In this study, we selected commonly used metrics for named entity recognition tasks, including accuracy (A), precision (P), recall (R), and F1 score, to evaluate the model's recognition performance. Accuracy indicates the proportion of all samples of the model that are predicted to be accurate. Precision

© STM Journals 2026. All Rights Reserved

refers to the proportion of true positive samples among the entities identified as positive by the model. Recall denotes the proportion of true positive samples that are correctly recognized. The F1 score is the harmonic mean of precision and recall, with the model performing optimally when the F1 score is 1 (indicating perfect precision and recall) and poorly when the F1 score is 0 [36, 37].

The overall model training process is shown in the following steps (Figure 5): Initially, all annotated data are divided into training, testing, and validation sets at a ratio of 8:1:1. Next, a joint extraction model is imported, and the training set is fed into the model for the initial round of training. After completing the training, the validation set is used to evaluate the model. The results are then analyzed to identify areas for optimization and guide further improvements.

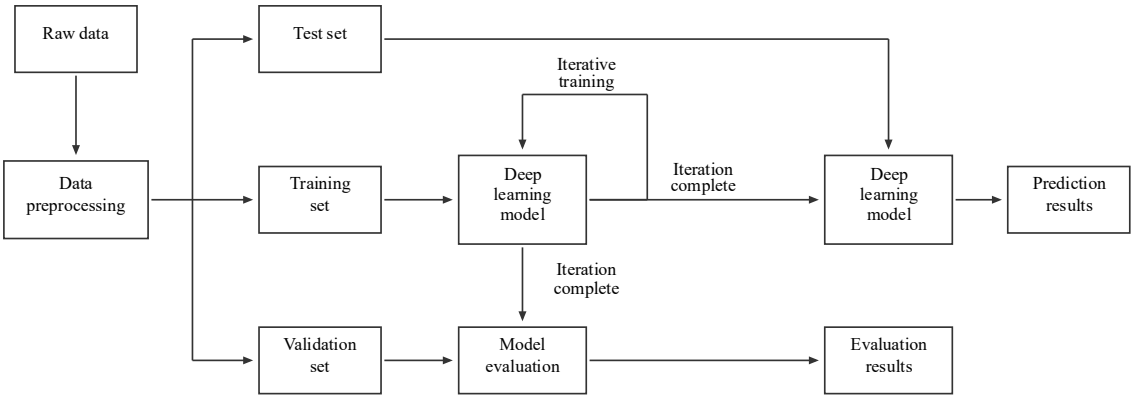


Figure 5. The overall process of model training.

The settings of the model parameters are shown in Table 3. The Adam optimizer was used during the model iteration process, and adversarial training was implemented using FGM (Fast Gradient Method).

Table 3. Model parameter settings.

Parameter name	Quantitative value
Maximum sentence length	256.
Batch size	8.
Rate of learning	2e-5.
Optimizer	Adam.
Function of loss	Cross Entropy Loss.

Knowledge Fusion and Processing

By calculating text similarity and standardizing vocabulary, synonyms were replaced with standard terms to achieve synonymy description and resolve entity ambiguities. Simultaneously, natural language processing techniques were used to identify coreference relations, and entity matching algorithms were employed to the consistency of the same entities across different reports. During the coreference resolution process, entities were supplemented and aligned, ensuring comprehensive and accurate representation. In this study, the entities and their attributes in the conceptual layer correspond one-to-one with those in the data layer, ensuring precise matching. Through a phase of knowledge fusion and processing, these knowledge elements were integrated into a hierarchical, large-scale knowledge system that facilitated unified management.

Construction and Application of the Knowledge Graph

Based on ontology construction, the acquired feature data was imported to establish the correspondence among entities, attributes, and relationships. Neo4j was then used to construct the knowledge graph. A usability study and evaluation of the breast cancer knowledge graph cloud service platform were conducted ground on the knowledge graph and electronic medical record information. This effectively integrated the dispersed breast cancer information into a standardized, systematic, and

© STM Journals 2026. All Rights Reserved

structured knowledge base. The breast cancer knowledge graph cloud service platform was designed and established using the SpringBoot framework, integrating electronic medical record information for enhanced functionality and efficiency. The design of the platform included the following main functional modules (Figure 6).

- *System Data Management:* Responsible for managing user information and permissions.
- *Knowledge Graph Management:* Provides structured management of the knowledge graph and basic data.
- *Knowledge Visualization Display:* Supports entity information management and graph-based visual querying.
- *Primary Applications of the Knowledge Graph Cloud Platform:* Includes knowledge graph-based visual analysis and intelligent retrieval, as well as intelligent question-answering and recommendation.

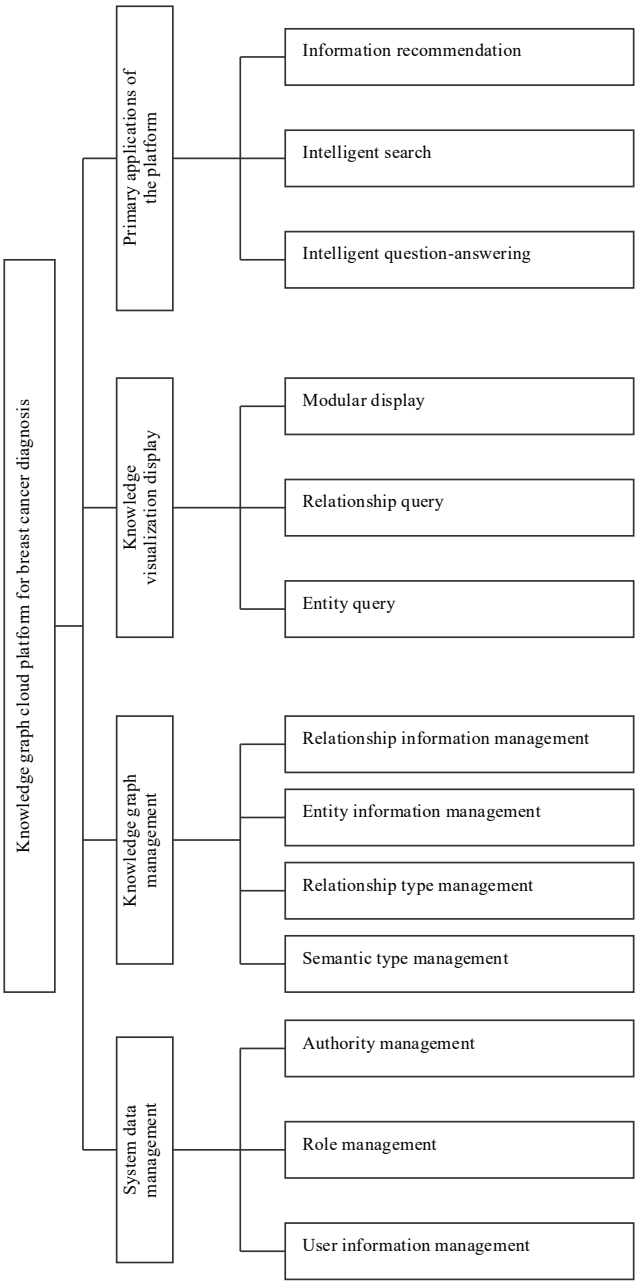


Figure 6. The function module of the platform.

© STM Journals 2026. All Rights Reserved

The platform architecture design adopted a Browser/Server (B/S) model and was developed using the SpringBoot framework. Its hierarchical structure primarily included three layers: data layer, service layer, and presentation layer (UI layer). The complete platform architecture is illustrated in Figure 7.

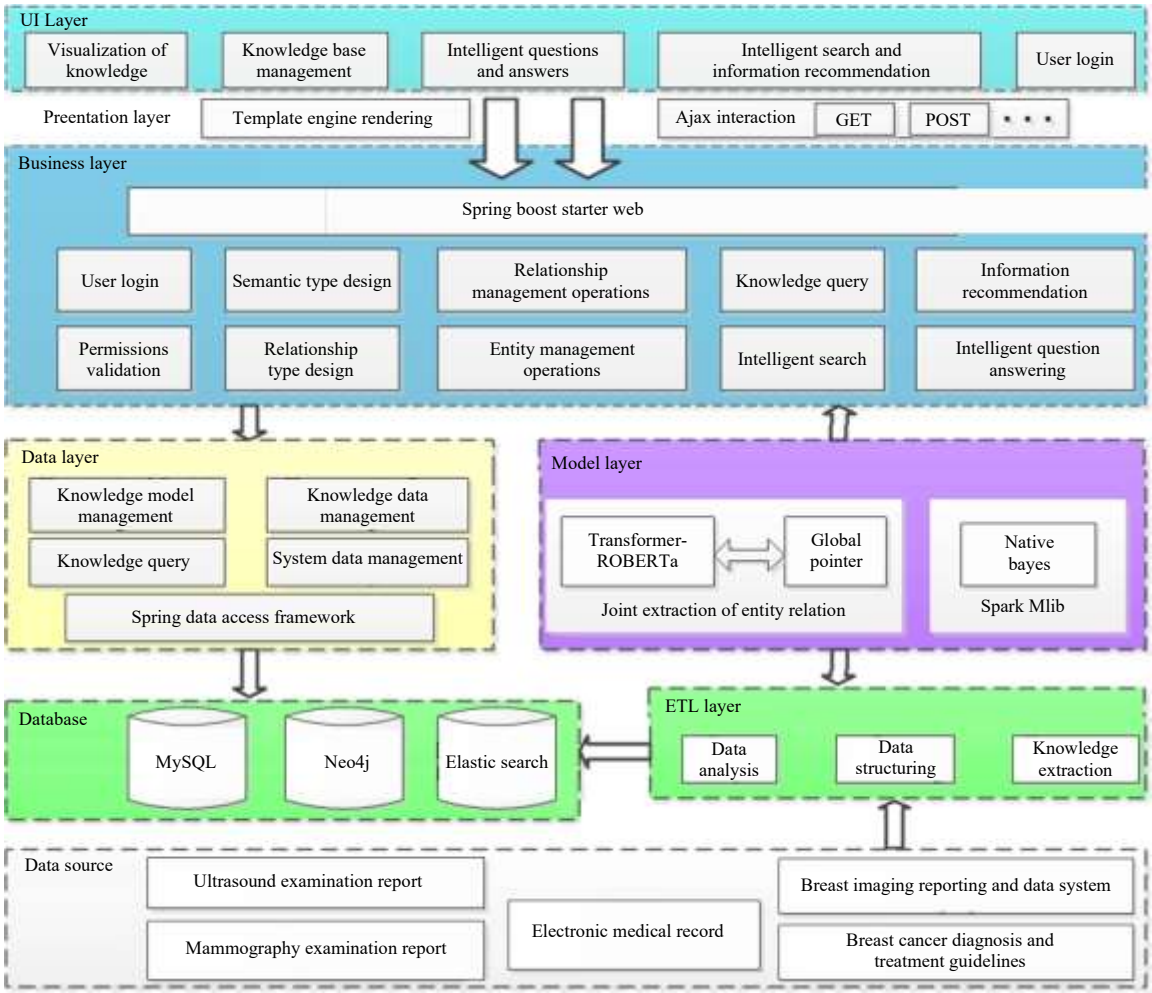


Figure 7. Platform architecture diagram.

The data layer serves as the core storage area. Neo4j, a high-performance NoSQL graph database, stores structured data in a network format rather than in tables. This approach makes Neo4j highly suitable for storing and representing entities, entity attributes, and relationships in a knowledge graph. MySQL is a secure, cross-platform, and efficient relational database with large storage capacity, used for managing user data and storing knowledge models. Elasticsearch is an open-source, distributed, RESTful-style search and data analysis engine used for handling plain text data with high retrieval efficiency. In this platform, Neo4j was used to store the entities and relationships within the breast cancer knowledge graph. The MySQL database managed the electronic medical record data and breast cancer knowledge models, while Elasticsearch indexed medical literature and reports to enable rapid retrieval. Data processing was carried out through ETL (Extract, Transform, Load) procedures to ensure data quality and structure.

The service layer, positioned between the presentation layer and the data layer, is responsible for data logic processing. Some functions required invoking services from the model layer, including the joint extraction of entities and relationships using the RoBERTa model combined with Global Pointer, to facilitate entity and relationship identification. Additionally, the system employed Spark MLib's Naive Bayes classifier for data classification and prediction, ensuring robust analytical capabilities. The Spring

© STM Journals 2026. All Rights Reserved

Data framework was utilized to maintain efficient interaction between the data layer and the service layer, allowing seamless data processing and retrieval.

The presentation layer facilitated data transmission through RESTful APIs, utilizing JSON data format for efficient front-end and back-end data exchange. On the front end, the Layui UI framework, built on HTML5 technology, was used for interface development. For data visualization, Echarts was employed to enhance user experience and data comprehension. This integration ensures smooth interactive data processing, providing an intuitive user interface.

RESULTS

Results of the Joint Extraction of Entities and Relationships

During the joint extraction process, variables from breast cancer ultrasound and mammography text were extracted, resulting in a dataset of 2000 entries. Table 4 categorizes and quantifies the entities within the breast cancer diagnostic knowledge graph. These entities categories cover various aspects of breast ultrasound and mammography, ranging from those with a larger number of entities, such as gland, side, and BI-RADS level, to those with fewer entities such as shape, surgical treatment, others, and prognosis. Most entity categories have fewer than 20 instances. Table 4 displays the focus of knowledge graph construction and data sources utilized.

Table 4. Distribution of entities.

Entity Category	Quantity	Entity Category	Quantity	Entity Category	Quantity
Gland	7680	Gland Typing	237	Sociology	42
Side	4589	Nipple	235	Location	38
BI-RADS Level	2054	Morphology Description	200	Other Treatments	32
Axilla	1834	Nipple Presentation	139	Gland Presentation	28
Lymphatic Presentation	1314	Size	129	Symptoms	27
Hyperplasia Presentation	1198	Clock-face Positioning	121	Location Description	19
Echo Intensity	1145	Duct	117	Shape	15
Lesion	669	Duct Presentation	83	Surgical Treatment	8
Quadrant	629	Edge Description	72	Other	1
Disease	392	Medication	66	Prognosis	1
Calcification Presentation	248	Examination	59		

According to Table 5, In terms of quantity, the three most significant relationships were gland-side, gland-BI-RADS level, and axilla-side, each with counts exceeding 1,000. This illustrates that in ultrasound and mammography examinations, the relationships between gland and side, gland and BI-RADS level, and axilla and side are very vital, likely to represent key indicators in breast examinations. Additionally, Table 5 also highlights more detailed relationships, such as those between characteristics and phenomena, including gland-hyperplasia performance, gland-echo intensity, and gland-calcification performance. Additionally, other aspects of the disease were also addressed, such as treatment and examination, encompassing “medication therapy,” “imaging examination,” “clinical manifestations,” “metastasis sites,” and “risk assessment factors.” Table 5 provides a comprehensive perspective, reflecting the diversity and complexity of the breast cancer knowledge graph.

Figure 8 shows that as FGM adversarial training progresses, the joint extraction experiences some oscillations during the initial learning phase. However, over time, the model’s accuracy gradually improves, recall increases, and both metrics stabilize with minimal fluctuations. In the later stages of training, the model’s overall performance tends to stabilize, indicating that it has achieved a balance between accuracy and completeness in entity recognition. This suggests that the model has reached its optimal performance given the current data and architecture.

© STM Journals 2026. All Rights Reserved

Table 5. Distribution of relationships.

Relationship category	Quantity	Relationship category	Quantity	Relationship category	Quantity
Gland-side	3070	Lesion-Margin description	72	Surgical treatment	8
Gland-BI-RADS grade	2054	Medical treatment	66	Adjuvant Therapy	7
Axilla-side	1341	Glands-Manifestations	28	Differential Diagnosis	6
Axillary-Lymphatic manifestations	1314	Imaging Studies	28	Lesion-Location description	6
Glandular-Hyperplastic manifestations	1198	Clinical picture	25	Ancillary tests	5
Gland-Echo intensity	1145	Site of metastasis	23	Related to (cause)	4
Gland-Quadrant	629	Chemotherapy	23	Histological examination	4
Glands-Lesions	446	Laboratory Tests	21	Genetic Factors	3
Gland-Calcified appearance	248	Complications	16	Related (symptoms)	3
Gland-Gland classification	237	Risk Assessment factors	16	Radiation therapy	2
Gland-Morphological description	200	Classification of pathology	16	Synonyms	2
Nipple-phenomenon	139	Site of disease	15	Symptoms after Treatment	2
Lesion-Size	129	Lesion-Shape	15	Prevention	1
Gland-Clock face	121	Gland-Location description	13	Screening	1
Nipple-Side	95	Risk Factors	12	Prognosis Survival rate	1
Duct-Performance	83	Related (transformation)	12	Nipple-Lesion	1
Duct-Side	83	Causes of disease	11		

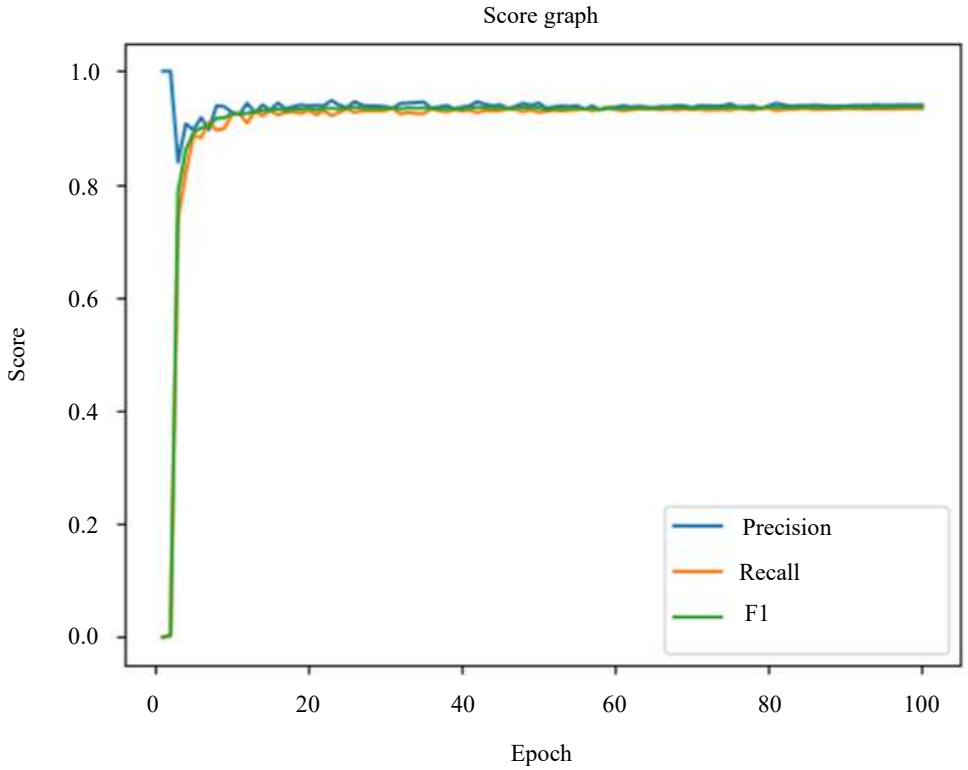


Figure 8. Line chart of training accuracy for the joint extraction model.

Table 6 presents the performance of different models in the joint extraction task. The proposed model in this research achieved an accuracy of 95.9%, a recall of 95.3%, and an F1 score of 95.6%. These metrics fully demonstrate the robust language understanding capabilities of the RoBERTa model and the advantages of the Global Pointer, demonstrating its effectiveness in this task.

© STM Journals 2026. All Rights Reserved

Table 6. Comparison of the performance of each model.

Model	Precision	Recall	F1
BiLSTM-CRF	92.8	90.56	91.15
BiLSTM-Attention-CRF	93.58	91.79	92.47
BiLSTM-Highway-CRF	93.61	91.77	92.52
Bert	94.0	93.3	93.7
RoBERTa-Global Pointer	95.9	95.3	95.6

Results of Knowledge Fusion

In this study, knowledge fusion was performed to address the issue of varying writing and descriptions of clinical terms related to "breast cancer." This involved semantic-based entity alignment, standardizing clinical terms to avoid information redundancy and misunderstandings caused by inconsistent naming of extracted medical entities. After aligning the extracted content, a total of 23,421 entities and 13,001 relationships were obtained. A comparison of some entities before and after fusion is shown in Table 7.

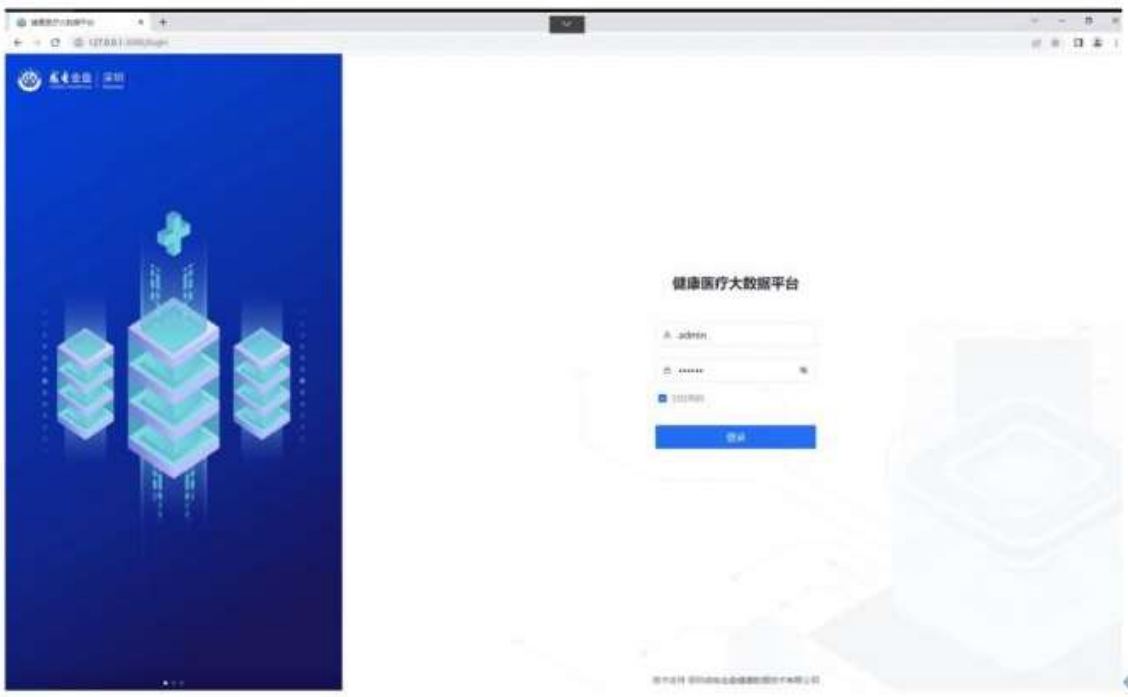
Table 7. Comparison of entities before and after fusion.

Original vocabulary	Standard vocabulary
BI-RADS:3	BI-RADS3 category.
BI-RADS 3 category	BI-RADS3 category.
BI-RADS III category	BI-RADS3 category.

APPLICATIONS OF THE KNOWLEDGE GRAPH CLOUD PLATFORM

Design and Implementation of the Breast Cancer Knowledge Graph Service Platform

After integrating electronic medical record information, a breast cancer knowledge graph cloud service platform has been designed and established using the SpringBoot framework. This platform effectively consolidated scattered clinical information into a systematized, standardized, and normalized knowledge database. The platform features a front–end interface, system management, data management, visualization interface, and intelligent retrieval interface, etc., as illustrated in Figure 9.



© STM Journals 2026. All Rights Reserved



Figure 9. Platform interface display.

© STM Journals 2026. All Rights Reserved

Knowledge Graph Visualization and Exploration

In the breast cancer knowledge graph, a node may represent breast cancer, while the connected nodes indicate various attributes and characteristics related to it such as tumor size, type, and location. The edges between these nodes reveal the relationships among these characteristics. Researchers do not need specialized medical professional knowledge. By entering specific query terms into the knowledge graph cloud service platform, it can obtain multi-feature information about patients such as basic information, examination information, classification information, etc. (Figure 10). This reduces time costs and improves work efficiency. The visualization of relationships provides doctors or researchers with an easily understandable and intuitive way to access information or discover new insights.

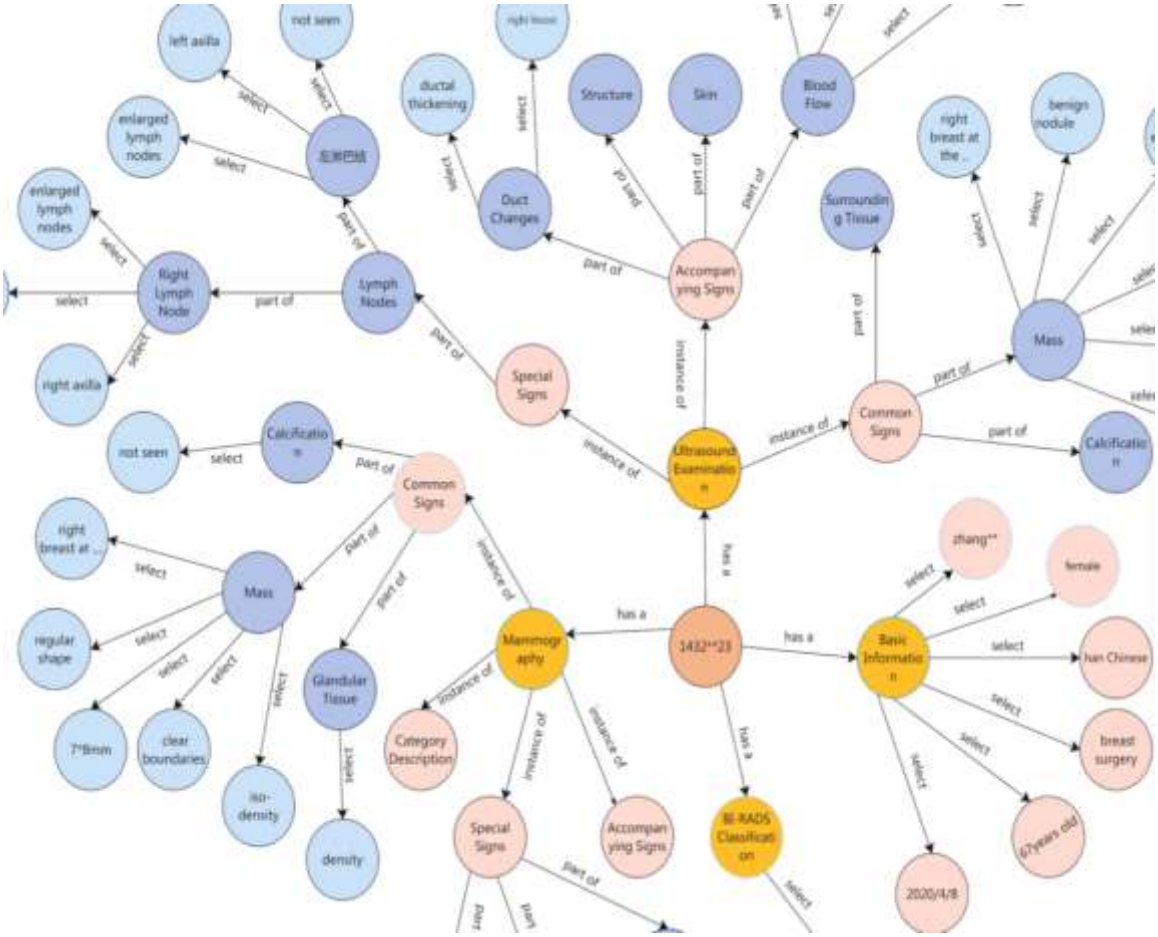


Figure 10. Examples of visualization.

Semantic Search and Query Interfere

Neo4j's "Search phrase" was utilized to customize query statements, including simple queries, combined queries, and conditional queries, enabling researchers to perform in-depth data access and analysis. For example, using predefined advanced query templates, clinicians can input query statements such as "patients aged between 50–52 years with a BI-RADS classification of 4a." This allows clinicians to retrieve basic patient information as shown in Figure 11.

Intelligent Diagnosis Assistance

Through intelligent assisted screening, some masses with regular shapes, clear boundaries, equal density, and uniform echoes were classified as BI-RADS 3a (Figure 12). A BI-RADS classification of 3a suggests that the mass is likely benign. Regular shape and clear boundaries can serve as indicators to rule out malignant tumors. Clinical assessment confirmed the accuracy of this conclusion. By further analysis, the impact of individual features or combinations of features on BI-RADS classification can

© STM Journals 2026. All Rights Reserved

be estimated. Utilizing intelligent search services to find the required diagnostic information can help clinicians obtain a clearer diagnostic perspective, thereby more effectively assisting in the diagnosis of breast cancer.

Efficient Knowledge Graph-Based Retrieval

This platform utilizes a knowledge graph-based intelligent retrieval method, supporting nested combination searches of multi-layered knowledge logic relationships. This approach shortens the retrieval cycle and enhances the efficiency and quality of data retrieval in clinical research. Here, an actual research project was used as an example to illustrate disease-specific intelligent retrieval. Clinical researchers first applied to join the data-sharing platform. The system administrator added these researchers to the system. Based on research projects, researchers could search for the required datasets, such as data on breast cancer patients, assessed as BI-RADS category 4. The system then explored the data according to the researchers' needs, utilizing the knowledge graph for precise data filtering. This allowed researchers to focus on analysis without worrying about the data source, providing more granular and accurate data. By selecting options related to "BI-RADS 4," specifically "BI-RADS 4a," "BI-RADS 4b," and "BI-RADS 4c," a total of 117 relevant data entries were retrieved. The system allowed users to export the relevant tables and fields according to their needs. These processes facilitated subsequent use such as exporting patient information and hospitalization records (Figure 13).

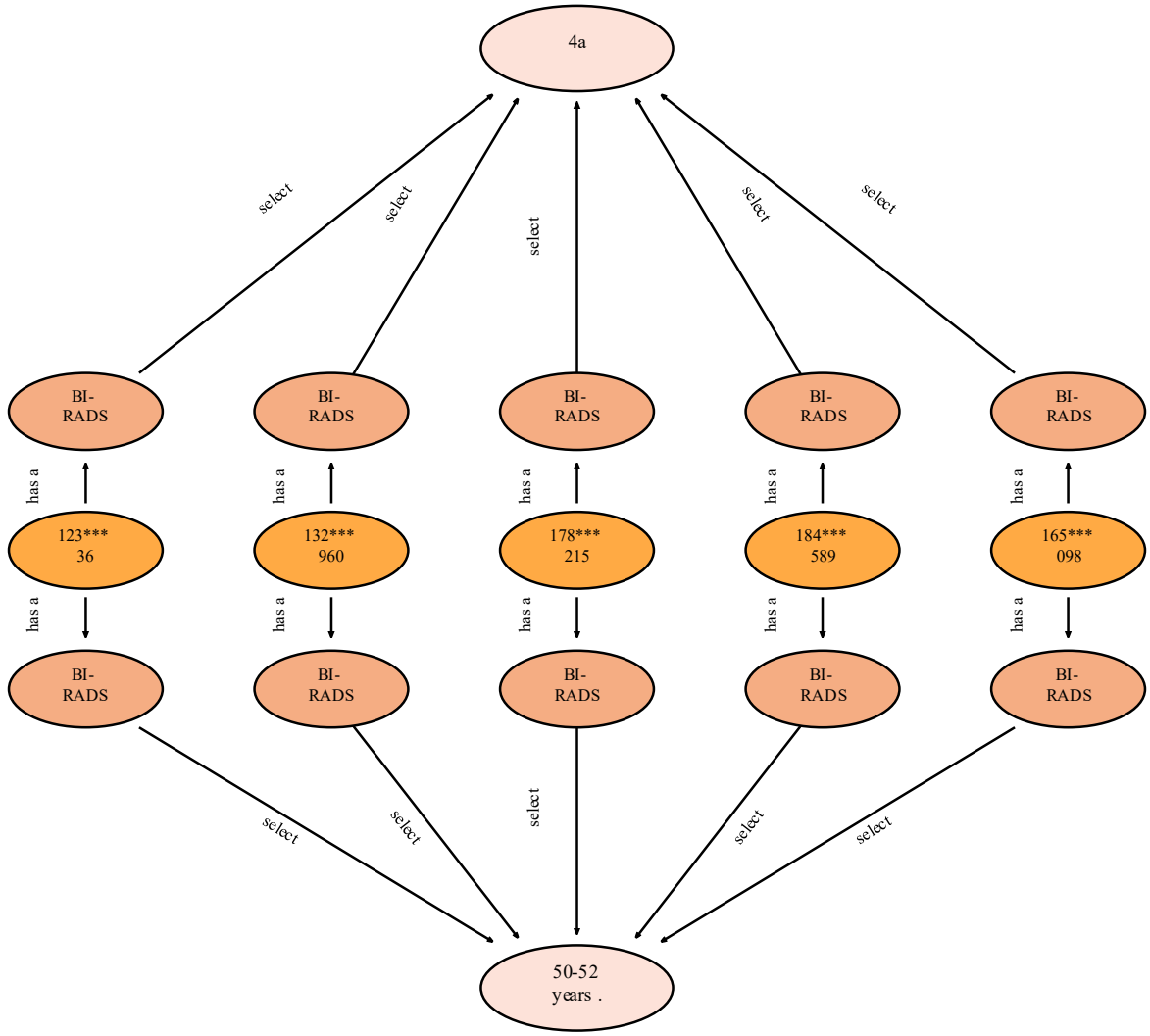


Figure 11. Query results for "information on patients aged between 50–52 with BI-RADS grade 4a".

© STM Journals 2026. All Rights Reserved

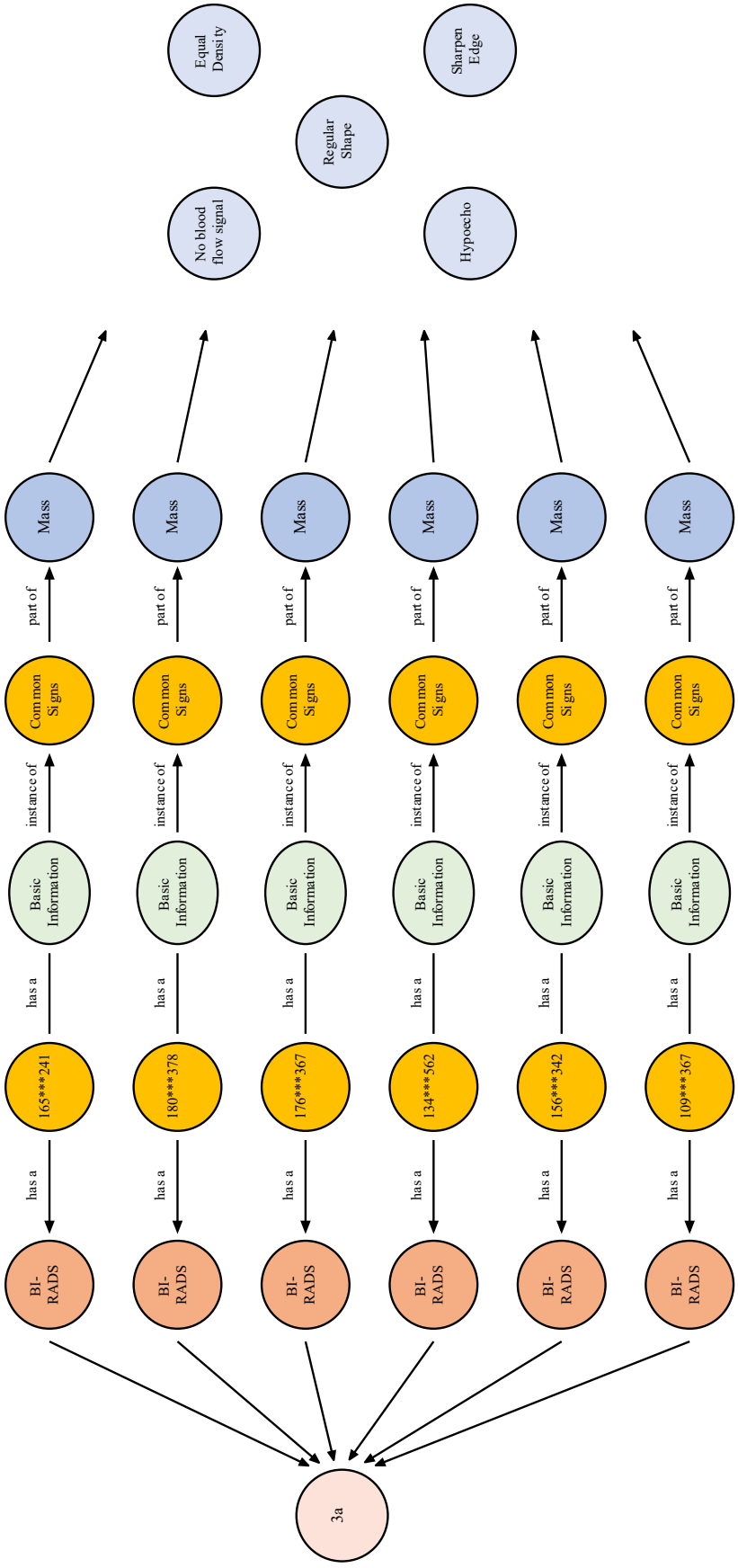


Figure 12. Corresponding relationship between entity features and BI-RADS classification.

© STM Journals 2026. All Rights Reserved





© STM Journals 2026. All Rights Reserved

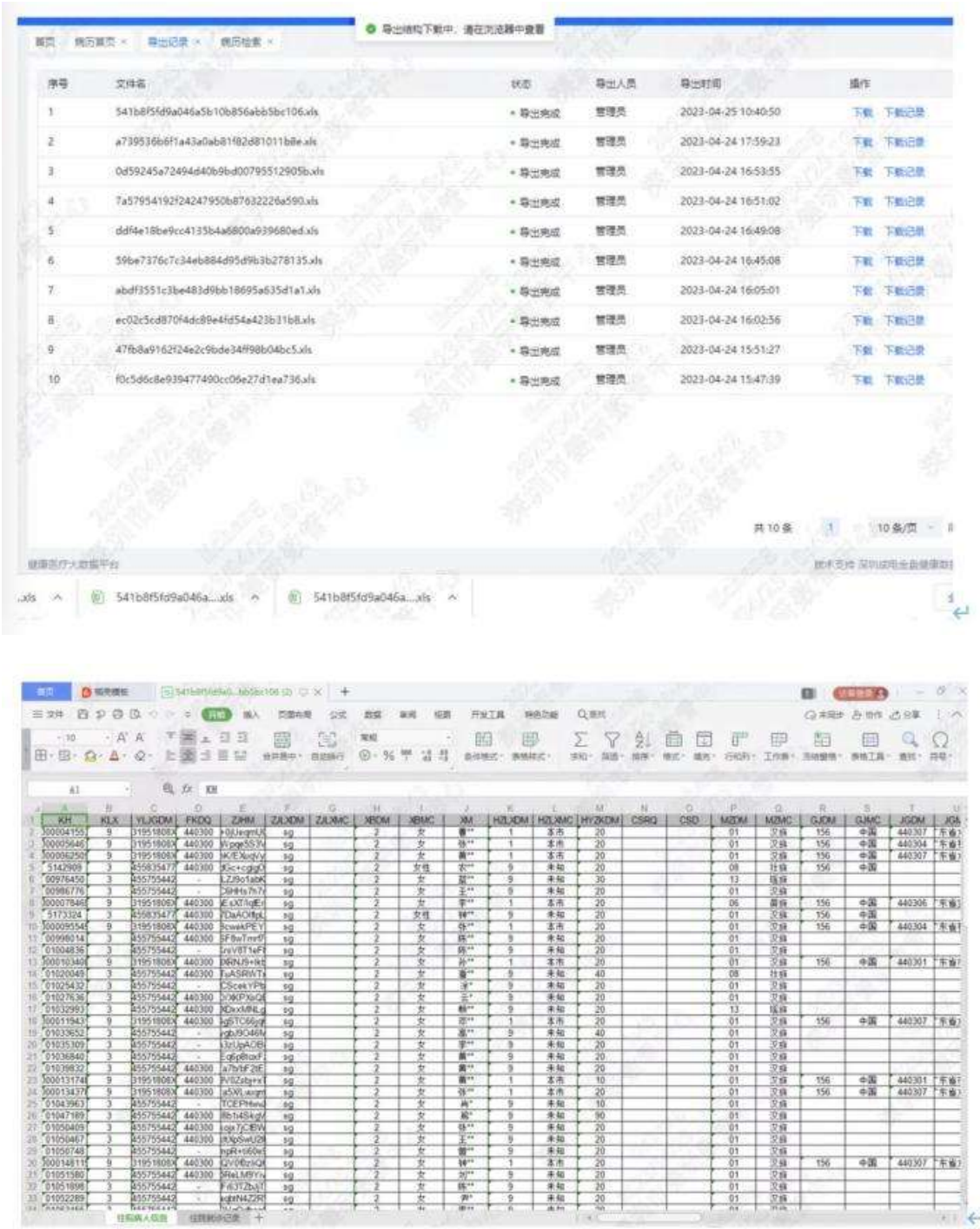


Figure 13. An example of intelligent retrieval.

Retrieval Efficiency

For vast medical information resources, an accurate retrieval strategy is crucial to meet users' needs for precise data retrieval. By comparing the speed and retrieval efficiency of our method with traditional retrieval methods, our approach shortens the retrieval cycle and improves the efficiency and quality of data retrieval in clinical research. Table 8 shows that at shallower depths; the retrieval speed of the knowledge graph is comparable to that of traditional retrieval methods. However, for queries beyond three levels, the retrieval speed of this knowledge graph significantly improves, surpassing traditional methods.

© STM Journals 2026. All Rights Reserved

Table 8. Comparison of retrieval speeds.

Depth of query	The average time of traditional retrieval /ms	The average time of knowledge graph retrieval /ms
1	71	85
2	89	98
3	139	103
4	184	107
5	238	114
>5	257	143

To verify the effectiveness of knowledge graph retrieval, the accuracy rate is used to evaluate the performance of the retrieval method. The formula for calculating accuracy is as follows:

$$\text{Accuracy Rate} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%$$

TP (True Positive), the number of texts retrieved that are relevant to the retrieval needs;

FP (False Positive), the number of texts retrieved that are not relevant to the retrieval needs.

The performance of two different retrieval methods on a Linux machine equipped with an Intel Core i7-7700K is shown in the table below (Figure 14). The result shows that the overall accuracy of our retrieval method is higher than that of traditional retrieval methods.

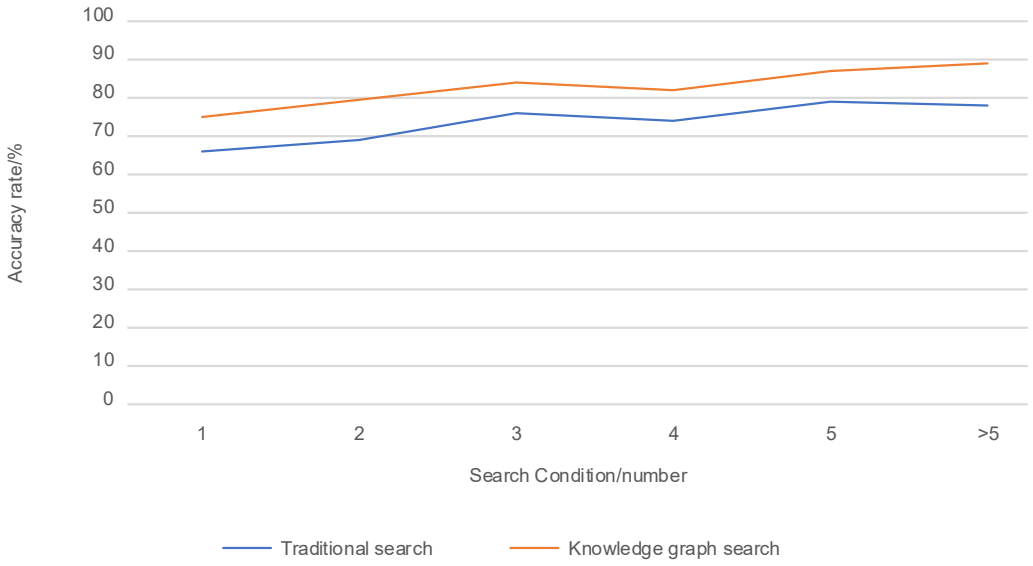


Figure 14. Comparison of the retrieval accuracy.

DISCUSSION

Breast cancer, is a prevalent disease affecting women's health, can significantly improve patient prognosis and survival rates through early diagnosis and treatment. Although breast ultrasound reports and mammography results are both important data sources for breast cancer diagnosis, many studies indicate that combining breast ultrasound with mammography is superior to using either method alone or can enhance diagnostic efficacy. We integrated the textual findings from ultrasound reports and mammography, incorporated not only the sonographic features about the tumor size, shape, and edges from ultrasound results but also mammographic attributes like tissue density and calcification patterns from mammography.

© STM Journals 2026. All Rights Reserved

The comprehensive descriptions of breast cancer diagnose within the nodes of the knowledge graph provide a holistic view, aiding users in quickly understanding the unique characteristics of breast cancer. This integrated approach enhances diagnostic sensitivity, enabling more informed and rapid diagnostic decisions. Previous studies have demonstrated that the integrated multiple imaging methods can improve diagnostic sensitivity, suggesting that combined techniques are superior to single methods. One study confirmed that artificial intelligence can significantly improve the efficiency and accuracy of breast cancer diagnosis by processing and analyzing complex mammography data. Our findings further demonstrate that the knowledge graph-based retrieval system significantly outperforms traditional search methods in accuracy, thereby offering more precise and reliable diagnostic support. The trained model enhances auxiliary diagnostics by improving the differentiation between benign and malignant tumors and enabling more refined BI-RADS classification assessments. This capability allows clinicians and researchers to efficiently elucidate relationships and distinctions across complex clinical concepts. Moreover, unlike previous efforts, our study incorporates electronic medical records alongside multimodal diagnostic data to construct a specialized breast cancer knowledge graph—an approach that remains underrepresented in existing literature. While such knowledge graphs exhibit broad coverage, but they often lack specialized depth. In contrast, our knowledge graph, which includes information from ultrasound and mammography, exhibits a high degree of specialization within its specific field, offering significant depth of knowledge. It is anticipated that future advancements will establish multimodal, breast cancer specific knowledge graphs as the mainstream paradigm. By combining various types of media information, including text, medical images or videos, and audio, these knowledge graphs will form network graphs that clearly display the hierarchical structure of entities and relationships related to breast cancer.

Currently, most medical knowledge graphs are constructed using a bottom–up approach, primarily relying on data from readily available sources such as medical literature, online repositories, and open databases. However, knowledge graphs built this way often fail to capture the complexities of real-world clinical practice, limiting their practical utility in supporting disease diagnosis and treatment planning. Consequently, the quality of these "bottom–up" graphs may not sufficiently meet clinical needs. The ontology construction of our knowledge graph utilizes an integrative approach, combining both top–down and bottom–up methods. The combined method integrates real-world diagnostic data, specific domain knowledge, and the expertise of clinical practitioners. Such a joint method of ontology building allows for the rapid collection of core concepts and relational information, significantly reducing the overall time quired for construction. Additionally, the incorporation of domain knowledge enriches the ontology with more detailed semantic layers, enhancing its utility and relevance in clinical contexts.

The project used a RoBERTa pre-trained model, built on the Transformer architecture, which can better understand contextual information and support transfer learning. This means that the model could utilize knowledge learned from large, general-domain corpora and apply it to specific domain tasks, thereby reducing the need for specialized domain data. Meanwhile, it can use pre-training to compensate for any shortcomings in annotated data. Traditional models require much more annotated data to achieve similar performance levels. However, we employed the RoBERTa model to get high accuracy with less annotated data compared to traditional models. Our results suggest that the RoBERTa model demonstrates high efficiency and stability in handling complex tasks with smaller datasets. The combined utilization of RoBERTa and Global Pointer provides powerful support for processing complex textual data that requires advanced entities and relationships recognition such as legal documents and medical records. The combination model offers a more precise and efficient solution for information extraction in these demanding fields, improving the ability to capture nuanced details and interconnections within the text. The combination model can effectively identify and link different parts of the dataset and handle the complexities between them, which is crucial for building a robust and informative knowledge graph. Based on the combination of RoBERTa and GlobalPointer, our data demonstrates that our model can effectively extract entities and relationships from medical texts, with performance significantly exceeding that of other advanced models. To deal with complex data, we devoted adversarial training to enhance the model's robustness. This ensures that the model remains effective even when faced with noisy or challenging data scenarios.

© STM Journals 2026. All Rights Reserved

This knowledge graph has a wide range of applications, with one of the most important being the implementation of a data-sharing platform. The platform supports various applications, including visual analysis, intelligent question-answering, intelligent retrieval, and disease-assistance decision-making. Its functionalities primarily encompass the following aspects: data acquisition and integration, data storage and management, data processing and analysis, knowledge graph construction, security and privacy protection, user interface and experience, clinical decision support, and data sharing and collaboration. Additionally, the platform offers advantages in cost reduction and continuous improvement and updates. Other breast cancer platforms may implement single-function aspects such as remote diagnosis, risk prediction and assessment, benign and malignant prediction, and intelligent diagnosis. Due to the multifunctionality of our platform, it has made significant contributions to improving the quality of breast cancer diagnosis and treatment data management. This study provides a scalable and reproducible workflow for building disease-specific knowledge graphs. Future work will incorporate multimodal data including imaging and genomics – and expand the model's applicability to other languages and clinical contexts. This research contributes a foundational step toward intelligent, data-driven healthcare systems tailored for precision medicine.

CONCLUSIONS

This study successfully designed, constructed, and applied a breast cancer-specific knowledge graph using multimodal clinical data from ultrasound and mammography reports. By introducing a joint entity-relationship extraction model based on RoBERTa and Global Pointer, we achieved high accuracy in extracting structured knowledge from unstructured medical texts. The resulting knowledge graph integrates entities and relationships into an intuitive and semantically rich framework, supporting advanced applications such as visual analysis, semantic querying, and diagnostic assistance. The implemented data-sharing platform effectively addresses challenges related to fragmented and heterogeneous medical data, enabling more efficient and accurate retrieval and analysis of breast cancer information. This system offers significant potential to enhance clinical decision-making, support multidisciplinary collaboration, and improve the overall quality of breast cancer care.

Author Contributions

XF and LH wrote the manuscript. XD modified the manuscript. All authors read and approved of the final manuscript.

Data Availability Statement

The data presented in this study are available in the article.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

DICOM Standard National and Local Collaborated Engineering Laboratory (KFKT2024002).

Ethics Approval Statement

This research was approved by the hospital ethics committee of Longgang Central Hospital approval number LCH 0024001.

REFERENCES

1. Acciavatti RJ, Lee SH, Reig B, Moy L, Conant EF, Kontos D, et al. Beyond breast density: Risk measures for breast cancer in multiple imaging modalities. Radiology. 2023;306(3). doi: 10.1148/radiol.222575.
2. Adoma AF, Henry NM, Chen W. Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP); 2020 Dec 18–20; Chengdu, China. Piscataway (NJ): IEEE; 2020. p. 117–121. doi: 10.1109/ICCWAMTIP51612.2020.9317348.

© STM Journals 2026. All Rights Reserved

3. Ahn JS, Shin S, Yang SA, Park EK, Kim KH, Cho SI, et al. Artificial intelligence in breast cancer diagnosis and personalized medicine. J Breast Cancer. 2023;26(5):405–418. doi: 10.4048/jbc.2023.26.e44.
4. An B. Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data. Math Biosci Eng. 2023;20(4):6776–6799. doi: 10.3934/mbe.2023314.
5. Aristokli N, Polycarpou I, Themistocleous S, Sophocleous D, Mamais I. Comparison of the diagnostic performance of magnetic resonance imaging, ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient history: A review. Radiography (Lond). 2022;28(3):848–856. doi: 10.1016/j.radi.2022.03.005.
6. Buchberger W, Geiger-Gritsch S, Knapp R, Gautsch K, Oberaigner W. Combined screening with mammography and ultrasound in a population-based screening program. Eur J Radiol. 2018;101:24–29. doi: 10.1016/j.ejrad.2018.01.022.
7. Cai Y, Zhang Y, Zhai C, Xu Z, Wang C, Xie J, et al. Research on construction method and application of knowledge graph for power transformer intelligent operation and maintenance. Authorea Preprints. 2023. doi: 10.22541/au.169140561.12345678.
8. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi: 10.1038/s41597-023-01960-3.
9. Cui X, Song C, Li D, Qu X, Long J, Yang Y, et al. RoBGP: A Chinese nested biomedical named entity recognition model based on RoBERTa and Global Pointer. Comput Mater Contin. 2024;78(3):3751–3768. doi: 10.32604/cmc.2024.046352.
10. Daowd A, Barrett M, Abidi S, Abidi SSR. Building a knowledge graph representing causal associations between risk factors and incidence of breast cancer. In: Public Health and Informatics. Stud Health Technol Inform. 2021;281:724–728. doi: 10.3233/SHTI210292.
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1. Minneapolis (MN): Association for Computational Linguistics; 2019. p. 4171–4186. doi: 10.18653/v1/N19-1423.
12. Feng J, Zhang R, Chen D, Shi L. A visualization method of knowledge graphs for the computation and comprehension of ultrasound reports. Biomimetics (Basel). 2023;8(8):560. doi: 10.3390/biomimetics8080560.
13. Goyal A, Gupta V, Kumar M. Recent named entity recognition and classification techniques: A systematic review. Comput Sci Rev. 2018;29:21–43. doi: 10.1016/j.cosrev.2018.06.001.
14. Guo J, Hu J, Zheng Y, Zhao S, Ma J. Artificial intelligence: Opportunities and challenges in the clinical applications of triple-negative breast cancer. Br J Cancer. 2023;128(12):2141–2149. doi: 10.1038/s41416-023-02241-8.
15. Halodová K. Vliv fyzioterapeutických přístupů v léčbě lymfedému u žen po ablaci prsu [thesis]. Brno (CZ): Masaryk University; 2025.
16. Jain H, Raj N, Mishra S. A sui generis QA approach using RoBERTa for adverse drug event identification. BMC Bioinformatics. 2021;22(Suppl 11):330. doi: 10.1186/s12859-021-04234-0.
17. Jiang G, Fan M, Li L. A cloud platform for remote diagnosis of breast cancer in mammography by fusion of machine and human intelligence. In: Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations. Proc SPIE. 2016;9789:97890E. doi: 10.1117/12.2217126.
18. Jin Y, Wang J, Jiang J, Song Y, Xi C, Qin K. Research on the construction and application of breast cancer-specific database system based on full data lifecycle. Front Public Health. 2021;9:712827. doi: 10.3389/fpubh.2021.712827.
19. Li D, Yang Y, Cui J, Meng X, Qu J, Jiang Z, et al. Joint extraction of Chinese medical entities and relations based on RoBERTa and single-module global pointer. BMC Med Inform Decis Mak. 2024;24(1):218. doi: 10.1186/s12911-024-02654-9.
20. Li T, Mello-Thoms C, Brennan PC. Descriptive epidemiology of breast cancer in China: Incidence, mortality, survival and prevalence. Breast Cancer Res Treat. 2016;159(3):395–406. doi: 10.1007/s10549-016-3947-0.
21. Li X, Sun S, Tang T, Lu J, Zhang L, Yin J, et al. Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: Development and usability study. BMC Med Inform Decis Mak. 2023;23(1):210. doi: 10.1186/s12911-023-02290-0.

© STM Journals 2026. All Rights Reserved

22. Liu H, Chen Y, Zhang Y, Wang L, Luo R, Wu H, et al. A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. Eur Radiol. 2021;31(8):5902–5912. doi: 10.1007/s00330-020-07630-5.
23. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv [cs.CL]. 2019. doi: 10.48550/arXiv.1907.11692.
24. Lyu H, Sha N, Qin S, Yan M, Xie Y, Wang R. Dynamic graph embedding for outlier detection. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Adv Neural Inf Process Syst. 2019;32.
25. Mann RM, Hooley R, Barr RG, Moy L. Novel approaches to screening for breast cancer. Radiology. 2020;297(2):266–285. doi: 10.1148/radiol.2020200172.
26. Nam KJ, Han BK, Ko ES, Choi JS, Ko EY, Jeong DW, et al. Comparison of full-field digital mammography and digital breast tomosynthesis in ultrasonography-detected breast cancers. Breast. 2015;24(5):649–655. doi: 10.1016/j.breast.2015.07.001.
27. National Breast Cancer Expert Committee, China Quality Control Center for Breast Cancer Prevention and Treatment. Guideline for the management pathway and quality control of breast cancer prevention and treatment in China's counties. Cancer Innov. 2025;4(3). doi: 10.1002/cai2.70005.
28. Ohnuki K, Tohno E, Tsunoda H, Uematsu T, Nakajima Y. Overall assessment system of combined mammography and ultrasound for breast cancer screening in Japan. Breast Cancer. 2021;28(2): 254–262. doi: 10.1007/s12282-020-01166-4.
29. Peng C, Xia F, Naseriparsa M, Osborne F. Knowledge graphs: Opportunities and challenges. Artif Intell Rev. 2023;56(11):13071–13102. doi: 10.1007/s10462-023-10465-9.
30. Ra M, Yoo D, No S, Shin J, Han C. The mixed ontology building methodology using database information. In: Proceedings of the International MultiConference of Engineers and Computer Scientists 2012. Hong Kong: Newswood Ltd.; 2012. p. 1085–1090.
31. American College of Radiology. Breast imaging reporting and data system (BI-RADS). 4th ed. Reston (VA): American College of Radiology; 2003.
32. Song C, Long J, Li D, Qu X, Lin F, Zhang X. RoBGP: A nested biomedical named entity recognition method based on RoBERTa and Global Pointer. In: 2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter); 2023 Jan 25–27; Honolulu, HI, USA. Piscataway (NJ): IEEE; 2023. p. 33–38. doi: 10.1109/SNPD-Winter57725.2023.10109124.
33. Wang H, Zu Q, Lu M, Chen R, Yang Z, Gao Y, et al. Application of medical knowledge graphs in cardiology and cardiovascular medicine: A brief literature review. Adv Ther. 2022;39(9):4052–4060. doi: 10.1007/s12325-022-02231-3.
34. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. J Med Libr Assoc. 2006;94(1):41–47.
35. Xue QL, Wang BX, Miao KH, Li XD, Yu Y, Li Z. Construction of knowledge base of Chinese medicine manufacturing. Zhongguo Zhong Yao Za Zhi. 2022;47(12):3402–3408. doi: 10.19540/j.cnki.cjcmm.20220119.101.
36. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60–66. doi: 10.1148/radiol.2019182716.
37. Yang Y, Lu Y, Yan W. A comprehensive review on knowledge graphs for complex diseases. Brief Bioinform. 2023;24(1). doi: 10.1093/bib/bbac543.

© STM Journals 2026. All Rights Reserved