Knowledge Graph–Enhanced Deep Learning for Investigating Covid-19 Pathogenesis

Deepthi Rani S. S.[1], Dr. Renu Aggarwal[2]

[1]Research Scholar Sunrise University, Alwar

[2]Assistant Professor (Research Supervisor), Sunrise University, Alwar

DOI: https://doi.org/10.51584/IJRIAS.2026.110400061

Received: 10 April 2026; Accepted: 16 April 2026; Published: 04 May 2026

ABSTRACT

Knowledge graphs (KGs), which represent entities and their relationships in a structured semantic network, have been widely applied in the study of various diseases such as thyroid disorders, cardiovascular diseases, and neurological conditions. However, current diagnostic approaches often face challenges including incomplete data integration, limited scalability, and reduced diagnostic accuracy. These limitations highlight the need for advanced methodologies capable of addressing the complex nature of COVID-19 diagnosis.The proposed research introduces a framework that integrates knowledge graphs with deep learning techniques for improved COVID-19 analysis. Initially, COVID-19 related datasets will be collected from publicly available repositories such as Kaggle, including information on viral characteristics, transmission patterns, clinical symptoms, and public health data. From these datasets, relevant entities and relationships will be extracted to construct a COVID-19-specific knowledge graph. The constructed KG will then be transformed into low-dimensional vector representations using embedding techniques, enabling semantic representation within a vector space.Based on this knowledge representation, a novel deep learning model will be developed to predict COVID-19 cases using relevant input features. The model will utilize virus-related word vectors and knowledge entity vectors derived from the knowledge graph. Through supervised learning, the model will be trained to classify samples based on COVID-19 related symptoms and associated features. The effectiveness of the proposed diagnostic model will be evaluated using standard performance metrics.By integrating knowledge graph construction with deep learning models, the proposed study aims to improve understanding of COVID-19 pathogenesis and support evidence-based decision making in pandemic management. This approach has the potential to provide an efficient and accurate diagnostic tool for the early detection and management of COVID-19 cases.

Keywords: knowledge graph, disease prediction, Electronic Medical Record, COVID-19, Deep Learning

INTRODUCTION

Knowledge graphs (KGs) provide a graph-based structure that enables the integration, management, and large-scale analysis of diverse information sources [1]. Compared with traditional relational databases or NoSQL models, graph-based approaches offer several advantages. They provide a clear and intuitive representation of complex relationships between entities, where nodes represent entities and edges represent interactions or associations [2]. This structure is particularly suitable for domains involving intricate relationships. Additionally, knowledge graphs support schema flexibility, allowing data structures to evolve over time and enabling more adaptable data management and analysis across various fields.

A knowledge graph represents real-world knowledge by organizing entities and their relationships in a machine-readable format [3]. The concept of the knowledge graph was popularized by Google in 2012 to enhance search engine capabilities. Typically, a knowledge graph stores information in the form of triples consisting of (entity, relation, entity), enabling the semantic representation of knowledge [4]. In such graphs, entities such as risk factors or medical conditions are identified through semantic analysis of textual data, while relationships between entities can be weighted to represent the strength of their associations. In addition to Google's Knowledge Graph,

several widely used open knowledge graphs exist, including DBpedia, Wikidata, ConceptNet, and Microsoft Concept Graph. These resources are multilingual and cover multiple disciplines.

In recent years, healthcare has experienced a significant transformation through the adoption of data-driven technologies for disease diagnosis, prognosis, and treatment. Within this evolving landscape, knowledge graphs have emerged as a powerful tool for representing complex biomedical information in a structured and semantically rich format. By modeling relationships among entities such as diseases, genes, proteins, drugs, and symptoms, knowledge graphs enable comprehensive data integration. This facilitates the discovery of hidden patterns, supports knowledge discovery, and assists clinicians and researchers in making informed decisions.

The outbreak of the COVID-19 pandemic in late 2019 created an unprecedented global health emergency, emphasizing the need for innovative methods to understand and manage infectious diseases [5]. Knowledge graphs have gained considerable attention in COVID-19 research due to their capability to model complex interdependencies among different aspects of the disease. By integrating data from diverse sources such as scientific literature, clinical records, genomic datasets, and epidemiological reports, knowledge graphs provide a comprehensive representation of the virus, its transmission patterns, clinical characteristics, and public health implications.

In parallel, numerous computational techniques have been developed for disease prediction. Traditional statistical approaches are commonly used in clinical decision-making, especially when known risk factors are available. Machine learning and data mining techniques, including both supervised and unsupervised learning methods, have also been widely applied to analyze case data and predict disease outcomes. Deep learning models such as convolutional neural networks (CNNs) and artificial neural networks (ANNs) have demonstrated strong performance in classification and risk prediction tasks. Other widely used predictive models include random forests, decision trees, and support vector machines.

Despite these advancements, the availability of Electronic Medical Record (EMR) data is often limited due to incomplete datasets and small sample sizes. Missing feature values can significantly affect the performance of predictive models, making accurate diagnosis and personalized disease prediction challenging. In this context, knowledge graphs offer a promising solution, as they can effectively integrate multi-source heterogeneous data, perform correlation analysis, and enhance disease prediction accuracy.

Consequently, knowledge graphs are increasingly being adopted in healthcare and medical research. By utilizing graph structures, these approaches capture complex relationships among diseases, symptoms, and risk factors more effectively than traditional statistical models. They are particularly valuable for analyzing synergistic effects among multiple variables. Building on this concept, the present research aims to develop a knowledge graph–based framework to model COVID-19 risk factors and their interactions, enabling improved understanding and predictive analysis of the disease.

LITERATURE REVIEW

Tiehua Zhou et al. (2024) [1] focused on lesion prediction in cervical cancer, using a large-scale knowledge graph constructed from medical literature and electronic medical record (EMR) data. They identified key risk factors through subgraph mining and proposed a lesion prediction algorithm. Their TRFLEX-LGP method outperformed existing models like BioBERT and WBC in high-quality phrase extraction. Through in-depth analysis, they established an ontology knowledge base and mined new key risk factors, enhancing prediction accuracy.

Rita T. Sousa and Heiko Paulheim (2024) [2] employed machine learning approaches to make predictions in order to address the global health challenge of diabetes, with a specific emphasis on the study of gene expression data. Through KGs, they presented a methodology that merged different gene expression datasets with domain-specific information. They converted patient data into vector representations for classifier input by using KG embedding techniques. The integration of heterogeneous datasets and domain expertise resulted in increased prediction performance, as demonstrated by the experimental results obtained from three GEO datasets relevant

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

to diabetes. The study emphasized the difficulties associated with small sample sizes in expression datasets and showed how well their method worked to get around this limitation.

The problem of creating precise knowledge graphs from pediatric Electronic Health Record (EHR) data was seized on by Mengyan Li et al. (2024) [3]. Using information from hierarchical medical oncology, pediatric EHR data, and general population EHR data, they proposed the Multi-source Graph Synthesis (MUGS) technique. MUGS effectively generated embedding for pediatric EHR codes, encompassing both uniform and non-uniform attributes across various healthcare facilities. They achieved adaptation to site-specific heterogeneity by carefully adjusting the hyper parameters. When it demonstrated robustness against negative transfer and improved effectiveness in identifying pediatric coding correlations, MUGS outperformed previous techniques. Their methodology makes it easier to do evidence-based research on pediatric populations and makes it possible to perform tasks like knowledge graph creation and phenotyping. One of the limitations is that the embedding dimensions have to remain consistent across sites.

In order to overcome the problem of scarce annotated datasets in the medical field, Majlinda Llugiqi et al. (2024) [9] integrated tabular data with KG embedding for the prediction of heart disease. Their work presented techniques for combining KGs with tabular data to improve the efficiency of machine learning (ML) algorithms. The study's methodology consisted of three steps: creating KGs, using embedding methods, and planning how to use embedding strategically. They achieved noticeable improvements by evaluating two embedding techniques across various machine learning models. They improved the F2 score for the K-Nearest Neighbors model from 71% to 80% and the accuracy of the Feed-Forward Neural Network from 82% to 85%. The effectiveness of using KG-based features was demonstrated by the results, which highlighted the significance of KG size and structure in ML performance enhancement. Although KG augmentation led to general performance improvements, the best ML model choice was still dependent on the properties of the dataset.

Knowledge Graph Embeddings (KGE) were investigated in the biological domain by Francesco Gualdi et al. (2024) [10] with the goal of representing complicated biological knowledge in a lower-dimensional space. In addition to developing a KG that integrated a variety of biomedical data, they also developed two new algorithms, DLemb and BioKG2Vec, to complement existing techniques. Experiments showed that their methods performed better in supervised and unsupervised situations. The study showed enrichment of disease-relevant activities among prioritized genes by using KGE to predict genes related with intervertebral disc degeneration (IDD). In order to maximize KGE production and predictive modeling, they also carried out a great deal of experimentation, including grid-search cross-validation.

Zhi-Qing Li et al. (2023) [6] focused on the prediction of diabetic macular edema (DME) in their study. They presented an AI-driven disease prediction model and emphasized the significance of early risk factor modification in lowering the incidence of DME. They addressed the problem of missing data in disease prediction by utilizing a knowledge graph, which improved speed and accuracy. With the use of correlation improvement techniques and statistical criteria, the model produced an 86.21% accuracy rate. 116 DME impacting factors were included in a medical KG that they created by carefully preparing the data and doing statistical analysis. By enabling early intervention and enabling tailored illness risk prediction, this method demonstrated the promise of clinical decision support systems in the management of disease.

Emmanuel Papadakis et al. (2023) [7] presented a comprehensive approach to constructing a KG for attention deficit hyperactivity disorder (ADHD). By integrating data from various sources including literature, clinical trials, medication information, and adverse effects, they automated the construction process. This KG aimed to facilitate in-depth exploration of adult ADHD, addressing the challenges posed by scattered knowledge. Through RDF conversion and information linking techniques, they successfully linked heterogeneous data sources, enabling seamless navigation and exploration. Evaluation through use cases demonstrated the KG's efficacy in enhancing information retrieval efficiency.

Jianchen Tang et al. (2023) [8] aimed to enhance recipe recommendation systems by considering time dynamics in user taste preferences. Using Knowledge Graph Attention Network (KGAT), they captured entity embeddings, integrating LSTM to predict users' future recipe preferences. This model, named PPKG, leveraged cancer knowledge graphs to offer personalized diet recommendations aiding in disease prevention and treatment. Their

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

contributions include incorporating time in recipe recommendations, introducing LSTM for sequence prediction, and extensive experimentation on self-created datasets, validating PPKG's effectiveness. The KGAT framework extracted entity embeddings, while LSTM mined connections in users' dietary records. Recipe recommendation was treated as a multiclassification problem, with LSTM output passed through a fully connected layer. Experimental results favored PPKG over baseline methods, showcasing its superior performance in recipe recommendation.

Huadong Xing et al. (2023) [9] developed the extensive Rare Disease Bridge (RDBridge) by employing text mining and knowledge graphs to provide a framework for acquiring data on rare diseases. They extracted entities and relationships from literature by utilizing deep learning models such as BioALBERT, and they achieved superior performance on a variety of models. RDBridge surpasses current databases in scope and depth by integrating gene, chemical, pathway, literature, and medical image data. The platform provides an intuitive user interface for accessing and viewing information about rare diseases. RDBridge is helpful in locating possible pathways and therapeutic possibilities, but further experimental confirmation is still required.

The intention of Dehai Zhang et al. (2022) [10] was to reduce the workload of radiologists by automating the generation of radiology reports from chest X-ray images. They integrated past medical information into their framework to overcome the shortcomings of the current deep learning techniques. They developed a knowledge graph to capture the interrelationships between diseases by mining correlations between medical findings. A graph neural network was trained with image-text hybrid characteristics that were taken from patient data to help aggregate previous knowledge for disease representation. On the Open-I and MIMICCXR datasets, the suggested strategy outperformed cutting-edge techniques in terms of report creation quality and clinical efficacy. The expressiveness of the model was further improved by utilizing structured labels from chest X-ray images.

Lulu Ding et al. (2022) [11] conducted a thorough investigation of ethical issues related to cerebral organoids using knowledge graphs and statistical analysis. They found that although these issues were discussed, the focus really intensified in 2017 as cerebral organoids became more like human brains and became involved in chimera research. Three types of ethical issues were identified via analysis: those that were common to the life sciences, those that were unique to brain organoids, and those that cut across several disciplines. These worries arose from advances in technology, particularly with regard to 3D culture systems and pluripotent stem cells. This proactive strategy seeks to address moral conundrums and promote the ethical advancement of brain organoids research in the field of biomedicine.

An automated approach combining text mining, knowledge graphs, and medical ontologies was developed by Michael Barrett et al. (2022) [12] to find mechanistic relationships between COVID-19 and chronic diseases such as diabetes mellitus (DM) and chronic kidney disease (CKD). Their method used SemRep to extract semantic relationships and added ontologies to PubMed articles starting in 2020. Using a KG to represent these correlations made it easier to analyze patterns and find answers to questions about how COVID-19 affects patients with DM and CKD. They discovered gene-disease connections and biological pathways by exploring the KG, which shed light on the course of the disease. Small sample sizes in several research and the labor-intensive manual process of creating visuals from the KG were challenges.

Gang Yu et al. (2022) [13] explored pediatric chronic disease management, highlighting challenges and resource misalignment in China. They proposed the Artificial Intelligence Chronic Management System (AICMS), integrating AI, knowledge graphs, big data, and IoT to optimize treatment and resource utilization. A classification model for asthma patients was developed using real healthcare data. The AICMS was designed to offer timely, active, and efficient chronic disease management for children, leveraging knowledge graphs to enhance operational efficiency. Through evaluation against the Chronic Care Model criteria, AICMS met user requirements, ensuring accurate information flow and intervention appropriateness. The system's structure enables hospitals, community doctors, and guardians to manage chronic diseases intelligently and efficiently, ultimately improving patient health and conserving resources.

Hegler C. Tissot and Lucas A. Pedebos (2021) [14] investigated miscarriage risk assessment using knowledge embedding techniques on clinical data. They explored various embedding strategies, demonstrating domain-

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

specific metadata's effectiveness in improving risk prediction accuracy. Utilizing a dataset of 24,877 pregnancies from the InfoSaude system, they represented categorical and numerical features as relations in a KG. Despite challenges like data sparsity and incomplete records, embedding methods enhanced machine learning applications for risk assessment. By analyzing the embedding neighborhood of each pregnancy, they optimized the risk score calculation, achieving the best F1 scores with a specific radius. Their approach prioritized explain ability and adaptability, distinguishing it from previous methods. The study showcased how embedding methods support comprehensive risk evaluation during pregnancy, offering insights into semantic correlations in clinical data.

By building a medical knowledge graph from the electronic medical records of patients with osteoarthritis in the knee, Xin Li et al. (2020) [18] enabled intelligent medical applications such as knowledge retrieval and decision assistance. They generated a domain ontology, used machine learning techniques to identify entities and extract relations, and then utilized a graph database to construct the knowledge graph. The research proved the dependability and thoroughness of the graph, confirming the efficiency of its development process. Using named entity identification and relationship extraction tools, they integrated patient data from different areas of electronic medical records to extract information. Their method solved challenges in locating entities in medical records and produced results with a high degree of accuracy.

Scope of the Study

Research on knowledge graph analysis for COVID-19 is highly significant due to its potential to enhance the understanding of the pandemic, support public health decision-making, improve clinical management, and contribute to vaccine development. As COVID-19 continues to evolve as a complex and rapidly changing global health challenge, innovative analytical approaches are essential to effectively address its diverse manifestations, transmission patterns, and societal impacts. Knowledge graph analysis enables the integration of heterogeneous data sources such as epidemiological records, clinical data, genomic information, and scientific literature, thereby providing comprehensive insights into the behavior and progression of the virus.

Effective public health policies, clinical interventions, and vaccination strategies rely on a clear understanding of viral biology, transmission dynamics, and patient outcomes. Knowledge graphs offer a robust framework for organizing and analyzing large volumes of interconnected data, allowing researchers to identify hidden patterns, detect risk factors, and predict disease spread. By capturing relationships among viral genomics, host characteristics, environmental influences, and clinical outcomes, knowledge graphs provide valuable insights into the mechanisms underlying COVID-19 pathogenesis and support the development of targeted interventions to reduce its public health impact.

Furthermore, knowledge graph analysis has broad applications in areas such as disease surveillance, outbreak prediction, diagnostic support, treatment optimization, and vaccine development. When combined with advanced deep learning techniques, these approaches enable the extraction of meaningful insights from vast COVID-19 datasets, improving strategies for pandemic monitoring and management. Integrating knowledge graph–based analytics into existing public health systems can also enhance data sharing, collaboration, and evidence-based decision-making at local, national, and global levels. Overall, the exploration of knowledge graph analysis represents a significant advancement in COVID-19 research, offering the potential to deepen our understanding of the virus and guide effective prevention, control, and treatment strategies.

PROPOSED METHODOLOGY

COVID-19-related datasets will be collected from Kaggle, a public repository known for its diverse collection of datasets. These datasets will include information from scientific literature, clinical records, genomic sequences, epidemiological data, and public health reports, covering various aspects of the virus, its transmission dynamics, clinical manifestations, and public health impact. A knowledge map of COVID-19 will be constructed by extracting entities and relationships related to the virus from the collected datasets. Entities such as viruses, hosts, symptoms, treatments, and epidemiological factors will be identified, and their relationships will be established based on the connections found in the data. A comprehensive knowledge map of the COVID-19 is

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IV April 2026

obtained in triple form, namely$< entity, relationship, entity >$. The constructed knowledge graph will provide a structured representation of COVID-19-related information. The detailed block diagram of the proposed work is shown in Figure 1.

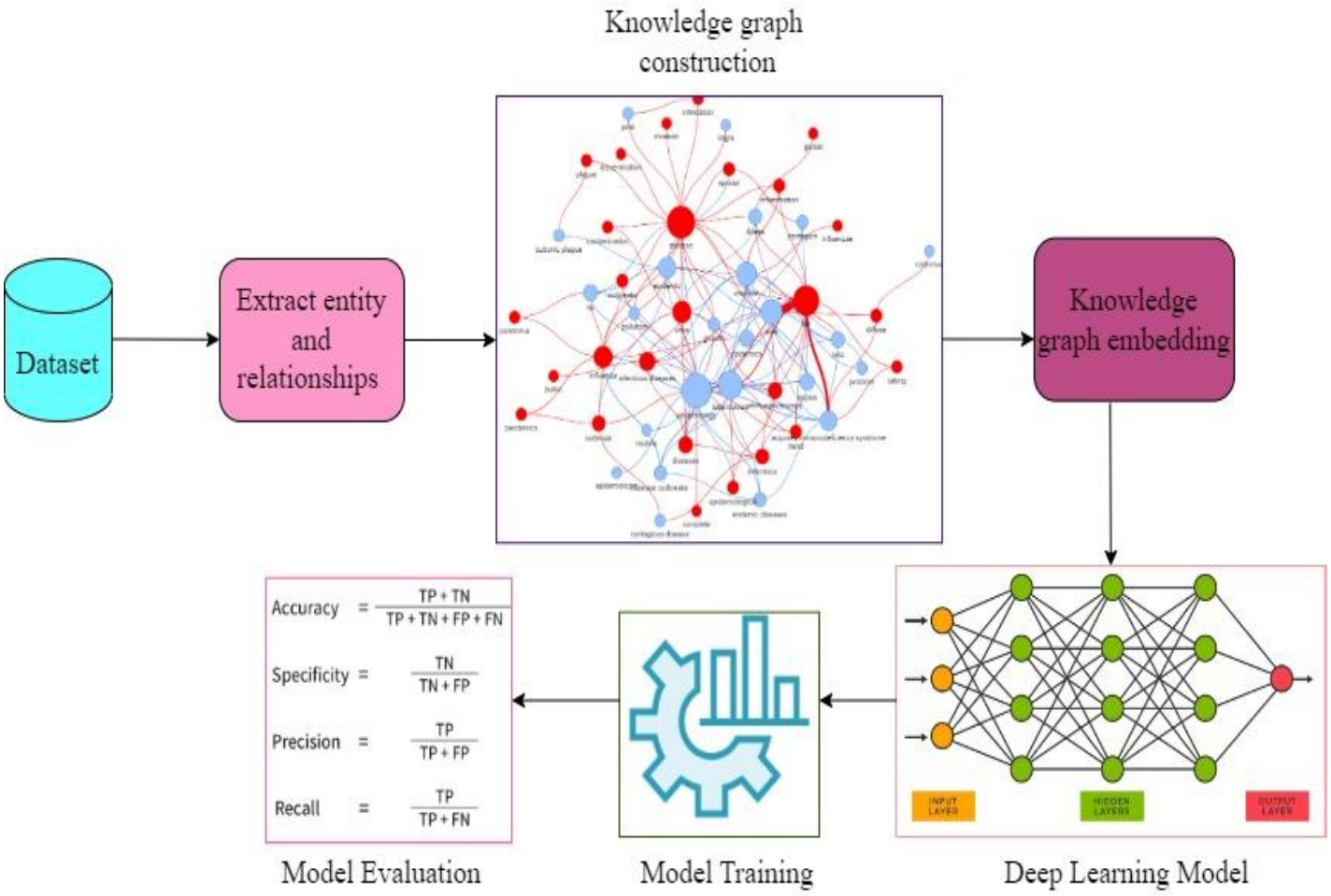

Fig.1: Block diagram of the proposed methodology

The entities and relationships in the constructed knowledge graph will undergo conversion into low-dimensional continuous vectors using knowledge graph embedding techniques. This step aims to capture the semantic meaning and contextual information of the entities and relationships, enabling them to be represented in a vector space. A novel Deep Learning Model will be trained using the constructed knowledge graph. The model will be designed to take as input the characteristic word vector of the virus and the relevant knowledge entity vector from the knowledge graph. Through supervised learning, the model will learn to predict whether the input sample exhibits symptoms of COVID-19. The performance of the trained diagnostic model will be evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1 score. The model will be tested on a separate dataset to assess its ability to accurately classify COVID-19 cases based on the input features.

CONCLUSION

In summary, the proposed methodology integrates knowledge graph construction, deep learning modeling, and performance evaluation to develop an effective and accurate diagnostic framework for COVID-19. By utilizing the rich relational information embedded within the knowledge graph along with the predictive capabilities of deep learning models, the approach aims to enhance the understanding of the virus and support evidence-based decision-making for pandemic management. Furthermore, the proposed framework is flexible and can be extended to the prediction and analysis of other diseases. As a direction for future research, the study also intends to explore the application of Graph Neural Networks (GNNs), which are specifically designed to operate on graph-structured data, as an alternative to traditional methods that rely on generating embeddings and using them as inputs for conventional machine learning models such as decision trees.

REFERENCES

1. Zhou, T., Xu, P., Wang, L., & Tang, Y. (2024). High-Risk HPV Cervical Lesion Potential Correlations Mining over Large-Scale Knowledge Graphs. Applied Sciences, 14(6), 2456.
2. Sousa, R. T., & Paulheim, H. (2024). Integrating Heterogeneous Gene Expression Data through Knowledge Graphs for Improving Diabetes Prediction. arXiv preprint arXiv:2404.14970
3. Li, M., Li, X., Pan, K., Geva, A., Yang, D., Sweet, S. M., ... & Cai, T. (2024). Multi-Source Graph Synthesis (MUGS) for Pediatric Knowledge Graphs from Electronic Health Records. medRxiv, 2024-01.
4. Llugiqi, M., Ekaputra, F. J., & Sabou, M. (2024). Leveraging Knowledge Graphs for Enhancing Machine Learning-based Heart Disease Prediction
5. Gualdi, F., Oliva, B., & Pinero, J. (2024). Predicting Gene Disease Associations With Knowledge Graph Embeddings For Diseases With Curtailed Information. bioRxiv, 2024-01.
6. Li, Z. Q., Fu, Z. X., Li, W. J., Fan, H., Li, S. N., Wang, X. M., & Zhou, P. (2023). Prediction of Diabetic Macular Edema Using Knowledge Graph. Diagnostics, 13(11), 1858.
7. Papadakis, E., Baryannis, G., Batsakis, S., Adamou, M., Huang, Z., & Antoniou, G. (2023). ADHD-KG: a knowledge graph of attention deficit hyperactivity disorder. Health information science and systems, 11(1), 52.
8. Tang, J., Huang, B., & Xie, M. (2023). Anticancer Recipe Recommendation Based on Cancer Dietary Knowledge Graph. European Journal of Cancer Care, 2023.
9. Xing, H., Zhang, D., Cai, P., Zhang, R., & Hu, Q. N. (2023). RDBridge: a knowledge graph of rare diseases based on large-scale text mining. Bioinformatics, 39(7), btad440.
10. Zhang, D., Ren, A., Liang, J., Liu, Q., Wang, H., & Ma, Y. (2022). Improving medical x-ray report generation by using knowledge graph. Applied Sciences, 12(21), 11111.
11. Ding, L., Xiao, Z., Gong, X., & Peng, Y. (2022). Knowledge graphs of ethical concerns of cerebral organoids. Cell Proliferation, 55(8), e13239.
12. Barrett, M., Abidi, S. S. R., Daowd, A., & Abidi, S. (2022). A Knowledge Graph of Mechanistic Associations Between COVID-19, Diabetes Mellitus, and Chronic Kidney Disease. Stud Health Technol Inform, 304-308.
13. Yu, G., Tabatabaei, M., Mezei, J., Zhong, Q., Chen, S., Li, Z., ... & Shu, Q. (2022). Improving chronic disease management for children with knowledge graphs and artificial intelligence. Expert Systems with Applications, 201, 117026.
14. Tissot, H. C., & Pedebos, L. A. (2021). Improving risk assessment of miscarriage during pregnancy with knowledge graph embeddings. Journal of Healthcare Informatics Research, 5(4), 359-381.
15. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37.
16. Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering, 34(1), 50-70.
17. Yuki, K., Fujiogi, M., & Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. *Clinical immunology*, *215*, 108427.
18. Li, X., Liu, H., Zhao, X., Zhang, G., & Xing, C. (2020). Automatic approach for constructing a knowledge graph of knee osteoarthritis in Chinese. Health information science and systems, 8, 1-8.
19. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. Queue, 17(2), 48-75.
20. Angles, R., & Gutierrez, C. (2008). Survey of graph database models. ACM Computing Surveys (CSUR), 40(1), 1-39.