ISSN 2393 - 9850
Pages : 33-46

A review on Insights into Covid-19 Pathogenesis: Exploring Knowledge Graphs and Deep Learning

Deepthi Rani S S
Research Scholar Sunrise University, Alwar

ABSTRACT

Knowledge graphs (KG), semantic networks representing entities and their relationships, have found widespread applications across various diseases, including thyroid disorders, cardiovascular diseases, and neurological conditions. However, existing methods in disease diagnosis suffer from limitations such as incomplete data integration, lack of scalability, and suboptimal diagnostic accuracy. These shortcomings underscore the need for innovative approaches to disease diagnosis that can address the unique complexities of COVID-19. The proposed research encompasses several key steps. Initially, COVID-19-related datasets will be gathered from repositories like Kaggle, covering diverse aspects including virus characteristics, transmission dynamics, clinical manifestations, and public health impact. Subsequently, a knowledge map specific to COVID-19 will be constructed by extracting pertinent entities and relationships from the collected datasets. This KG will then undergo conversion into low-dimensional continuous vectors using embedding techniques, facilitating semantic representation in a vector space. A novel Deep Learning Model will be developed using the constructed knowledge graph, with the aim of predicting COVID-19 cases based on input features. This model will intake the virus's characteristic word vector and relevant knowledge entity vector from the knowledge graph. Through supervised learning, the model will be trained to classify input samples indicative of COVID-19 symptoms. Finally, the performance of the trained diagnostic model will be assessed using standard metrics to gauge its effectiveness in diagnosing COVID-19 cases accurately. The proposed study holds immense potential in enhancing our comprehension of COVID-19 and aiding evidence-based decision-making in pandemic management. By amalgamating KG construction with deep learning modeling, this approach seeks to furnish an efficient and accurate diagnostic tool, enabling prompt identification and management of COVID-19 cases. Ultimately, this research aims to contribute significantly to

Received : 07 October 2024
Revised : 17 October 2024
Accepted : 21 November 2024

global endeavors in curbing the spread of the virus.

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

1. INTRODUCTION

Knowledge graphs with their graph-based structure make it easier to integrate, manage, and extract insights on a wide scale from a variety of information [1]. Graphs have several benefits over relational models or NoSQL substitutes. For a variety of domains, where edges and pathways express intricate interactions between entities, they offer a clear and understandable representation that is appropriate [2]. Additionally, schema flexibility is made possible by graphs, allowing data to change over time. Because of their ability to define schemas with flexibility knowledge graphs facilitate more flexible data administration and analysis and making them an effective tool for exploring complicated information across several fields.

KG represents the objective world by structuring entities and their relationships in an entirely machine-readable format [3]. In an effort to boost search engine performance, Google first introduced the idea of a knowledge graph in 2012. A knowledge graph serves as a repository containing diverse entities, concepts, and their semantic connections, represented in triple format (entity, relation, entity) [4]. Risk factors being entities require identification through semantic analysis of text. Weights are used to measure the relationships between entities; closer relationships are indicated by larger weights. Much other open KGs, including DBpedia, Wiki data, Concept Net, and Microsoft Concept Graph, are also extensively used globally, in addition to Google's Knowledge Graph. These resources are multilingual and multidisciplinary. In recent years the field of healthcare has witnessed a paradigm shift towards the integration of data-driven approaches for disease diagnosis, prognosis, and treatment. Amidst this transformation, knowledge graphs have emerged as a powerful tool for representing and organizing complex biomedical information in a structured and semantically rich manner. By encoding relationships between entities such as diseases, genes, proteins, and drugs, knowledge graphs facilitate comprehensive data integration, enabling researchers and clinicians to uncover hidden patterns, identify novel insights, and make informed decisions.

The outbreak of the COVID-19 pandemic in late 2019 presented an unprecedented global health crisis, underscoring the urgent need for innovative approaches to understanding and combating infectious diseases [5]. Knowledge graphs, with their ability to capture and model complex inter dependencies between various facets of a disease, have garnered significant attention in the context of COVID-19 research. Leveraging diverse data sources such as scientific literature, clinical records, genomic sequences, and epidemiological data, knowledge graphs offer a holistic view of the virus, its transmission dynamics, clinical manifestations, and public health impact.

Many techniques for disease prediction have emerged in recent years. Statistical methods, which are frequently applied in clinical decision-making, provide descriptive analysis for infectious diseases, particularly in cases where risk variables are well-known. In order to predict diseases using case data, machine learning and data mining—including supervised and unsupervised algorithms—have gained popularity. Deep learning models like convolutional and artificial neural networks perform exceptionally well in classification as well as risk prediction tasks. Other popular techniques for modeling and forecasting risks include random forests, decision trees, and support vector machines. Nevertheless, incomplete data and small sample sizes make it difficult to gather Electronic Medical Record (EMR) information, which makes identifying critical risk factors necessary. But when feature values are missing, these models can have trouble diagnosing diseases and fulfilling the demands of individualized disease prediction. A knowledge graph has emerged as a potential solution to these problems since it can perform correlation analysis, analyze multi-source heterogeneous data effectively, and forecast disease outcomes with accuracy. As a result, a knowledge graph is becoming more and more common in the healthcare and medical industries. These algorithms use graph representations to capture complicated interactions between diseases, symptoms, and risk factors. They are more accurate than traditional statistical methods and are especially useful in analyzing synergistic effects. This research will develop a knowledge base of risk factors and their interactions related to COVID-19 by utilizing graph structures.

2. 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 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 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-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.

3. SCOPE OF THE STUDY

The research of knowledge graph analysis for COVID-19 holds significant relevance due to its potential to advance the understanding of the pandemic's dynamics, inform public health strategies, improve clinical management, and support vaccine development efforts. It is a critical area of research with wide-ranging implications for global health and healthcare systems. The key and most significant importance of this study resides in its potential to aid in the continuous endeavors to manage the COVID-19 pandemic. COVID-19 presents itself as a multifaceted and swiftly changing public health challenge, with various expressions, modes of transmission, and societal repercussions. By leveraging knowledge graph analysis, researchers and healthcare professionals can integrate heterogeneous data sources, including epidemiological data, clinical records, genomic sequences, and scientific literature, to gain insights into the virus's behavior, identify emerging trends, and inform evidence-based interventions.

The success of public health measures, clinical interventions, and vaccination campaigns depends on a thorough understanding of the virus's biology, transmission patterns, and clinical outcomes. Knowledge graph analysis offers a powerful framework for synthesizing and analyzing vast amounts of data, enabling researchers to uncover hidden patterns, identify risk factors, and predict disease spread. By elucidating the complex relationships between viral genomics, host factors, environmental variables, and disease outcomes, knowledge graphs can provide valuable insights into the mechanisms underlying COVID-19 pathogenesis and inform targeted interventions to mitigate its impact on public health.

Moreover, the analysis of knowledge graphs holds promise for a broad spectrum of applications, encompassing disease surveillance, outbreak prediction, diagnostic aid, treatment refinement, and vaccine advancement. Through the utilization of sophisticated methodologies in

Deep Learning, valuable insights can be gleaned from the extensive pool of COVID-19 data accessible, thereby bolstering the efficacy of strategies aimed at pandemic mitigation and management. Additionally, the incorporation of knowledge graph analysis into established public health frameworks stands to bolster data exchange, cooperation, and decision-making processes across local, national, and global spheres. This collaborative effort is positioned to promote more synchronized and effective responses to the pandemic, nurturing a collective effort in the ongoing management of COVID-19.The exploration of knowledge graph analysis for COVID-19 will marks a crucial advancement in the response to the pandemic, offering the potential to transform the comprehension of the virus and guide evidence-based approaches to prevention, control, and treatment.

4. 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 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.

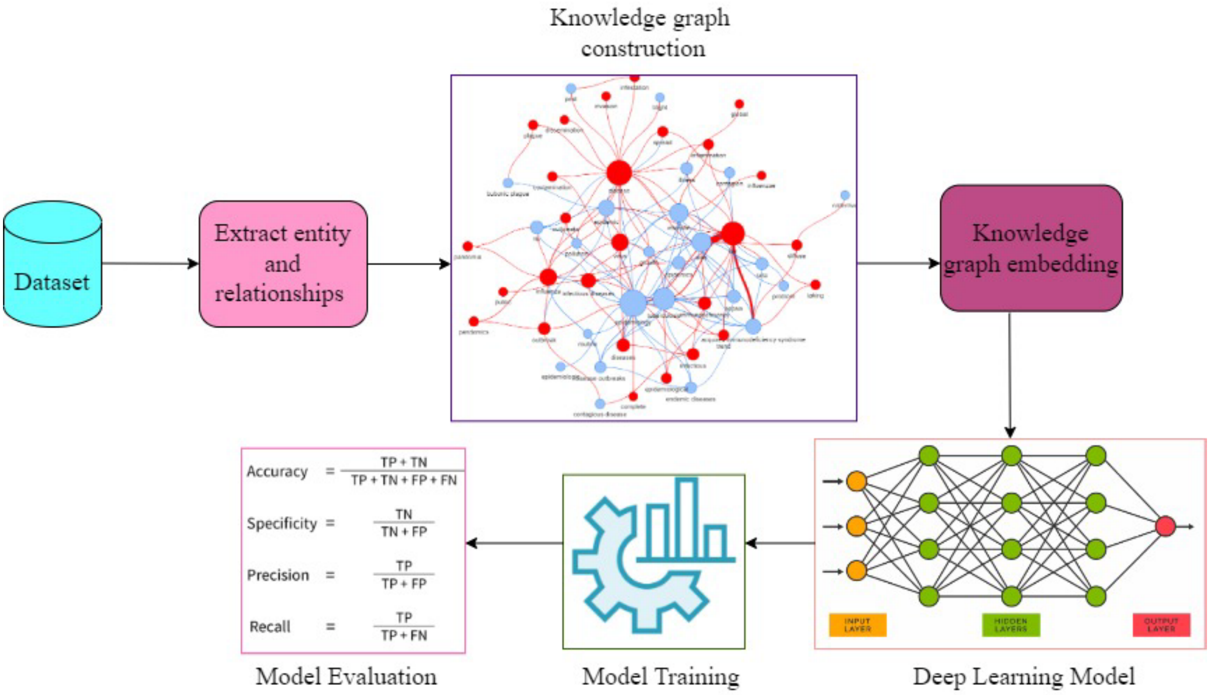
Knowledge graph construction
Dataset
Extract entity and relationships
Knowledge graph embedding
Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$
Specificity $= \frac{TN}{TN + FP}$
Precision $= \frac{TP}{TP + FP}$
Recall $= \frac{TP}{TP + FN}$
INPUT LAYER
HIDDEN LAYERS
OUTPUT LAYER
Model Evaluation
Model Training
Deep Learning Model

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.

5. CONCLUSION

In summary, the proposed methodology will aim to integrate knowledge graph construction, deep learning modeling, and performance evaluation to develop an efficient and accurate diagnostic tool for COVID-19. By leveraging the rich information encoded in the knowledge graph and the predictive power of deep learning models, the proposed approach will seek to improve our understanding of the virus and support evidence-based decision-making in the management of the pandemic. The proposed approach is versatile and can be extended to the prediction of other diseases. In addition, since graph neural networks has gained substantial traction recently, as future work, we aim to investigate how can these architectures explicitly designed for graph structures can be used rather than the conventional process of generating embeddings and given them as input for classical machine learning methods 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.