Journal of Computational Medicine and Informatics
Vol. 1, No. 2, Apr - Jun 2025, pp. 30-35
ISSN: 3107-3735 (O), DOI: https://doi.org/10.17051/JCMI/01.02.04

Graph Neural Network-Based Modeling of Disease Progression Using Longitudinal Clinical Data

Sanjay P Dhangar

Clinical HOD & Assistant Professor, Department of Urology, Bharati Hospital & Research Centre, BVDTU, Pune, Maharashtra, India, Email: drspdhangar@gmail.com, Orchid id - 0000-0002-0932-9414

Article Info

Article history:

Received : 12.04.2025
Revised : 11.05.2025
Accepted : 08.06.2025

Keywords:

Graph Neural Networks,
Disease Progression,
Longitudinal Data,
Clinical Informatics,
Temporal Modeling,
Healthcare AI

ABSTRACT

Simulation of the clinical dynamics of disease progression based on longitudinal clinical data is critical to support early diagnosis, stratification of risks, and individual planning of therapeutic approaches in contemporary healthcare systems. But, more traditional machine learning and sequence-based models typically have difficulty jointly modeling intricate temporal and inter-feature relationships found in electronic health records. In an effort to manage this weakness, this paper introduces a Graph Neural Network (GNN)-based disease progression modeling of longitudinal clinical data. Patient records in the proposed approach are modeled as dynamic graphs, with the nodes being the clinical variables, and the edges containing physiological correlations and time-dependent interactions among the clinical variables on the same time steps. The model combines graph convolutional and temporal aggregation to acquire expressive disease evolution representations. The MIMIC-III clinical data, which consists of a wide range of patient trajectories, and a diverse set of features are experimentally evaluated. The suggested approach outperforms the baseline models, such as Logistic Regression, Random Forest, and Long Short-Term Memory (LSTM) networks with a difference of up to 57% in the prediction accuracy and ROC-AUC. Moreover, the framework offers a better understanding of results, revealing the clinically relevant interactions of features. The findings indicate that the GNN-based methods are effective at representing both structural and temporal aspects of healthcare data, which can become an effective solution to advanced clinical decision support systems.

1. INTRODUCTION

The modern healthcare systems have integrated disease progression modeling as an important aspect especially with regards to chronic and degenerative diseases like cardiovascular diseases, diabetes, and neurodegenerative diseases. Proper disease trajectory modeling allows efficient disease diagnosis, risk assessment, patient-specific treatment planning, and, as a result, patient outcomes and cost reduction in healthcare. The growing access to longitudinal clinical information, such as electronic health records (EHRs), laboratory measurements, and medical imaging, comes with unprecedented chances to examine how health in patients changes over time. However, all these data are heterogeneous, high-dimensional, sparse, and irregularly time-varying, which makes them very difficult to analyze using traditional methods.

Conventional statistical models, including Cox regression and hidden Markov models, are easy to interpret yet do not fit any complex nonlinear relationships. Deep learning models, such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, recently have been used extensively in healthcare to support time modeling [1], [2]. Although these models are useful at modeling sequential dependencies, they generally model clinical variables on their own and disregard the underlying relational structure of features, patients, and states. The graph-based learning has become an effective paradigm towards modeling complex relationships in the structured data. Graph Neural Networks (GNNs), such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), have produced good performance in healthcare tasks, like patient similarity modeling and disease prediction [3]-[5]. All of the existing literature is however based on graphic representations and does not successfully incorporate temporal dynamics inherent in longitudinal clinical data. This makes them unable to precisely simulate disease evolution over time.

To overcome these limitations, this paper offers a Graph Neural Networks-based design of the disease progression model based on longitudinal clinical data. The approach proposed describes patient data as dynamic graphs, which allows the incorporation of both time varying and relational information. The model learns the meaningful representations of disease trajectories by integrating graph convolution operations with a set of temporal aggregation mechanisms.

The key findings of this paper can be summarised as follows:

1. Longitudinal clinical data in the form of a graph, revealing both the relationship between features and temporal relationships.
2. A learnable GNN-based modeling mechanism to learn the patterns of disease progression.
3. Elaborate analysis showing better predictive performance than conventional and deep learning baselines.

The rest of the paper is structured in the following manner: Section 2 reviews related work, Section 3 provides the proposed methodology, Section 4 explains the experimental setup, Section 5 discusses the results and Section 6 provides a conclusion of the paper.

2. RELATED WORK

Statistical, machine and deep learning methods have been widely used to model disease progression. Initial approaches were mainly based on survival analysis models like Cox proportional hazards models and hidden Markov models (HMMs) that gave interpretable information on the dynamics of diseases [6], [7]. Nevertheless, they are not as effective in the description of nonlinear relationships and interactions among clinical variables. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are among the most popular sequential models that are commonly applied to longitudinal healthcare data with the development of deep learning technologies [8], [9]. The models are very effective in capturing the temporal dependencies, and have been effectively used in tasks like predicting disease onset and predicting patient outcomes. Although they have strengths, RNN-based approaches generally assume clinical features as independent inputs, and do not capture the underlying relational relationship between variables and the patients. To overcome this drawback, the use of graph-based learning algorithms has been receiving rising interest in the recent years. Graph Neural Networks (GNNs or Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs)) are a potentially useful tool to model structured relations on healthcare data [10], [11]. These have been used to simulate patient similarity networks, disease classification and prediction of drug interactions where they have shown better results than the traditional ones. More recent works have tried to adopt temporal information in graph-based models. Other strategies like time-evolving GNN and dynamics graph learning models have demonstrated potential in tracking dynamic relationship changes [12] - [14]. Nonetheless, most literature uses either fixed graph models, or represents time dynamically, often omitting the concomitant modelling of temporal dynamics and high-order feature interactions [15] - [17].

Also, issues like the sparsity of data, non-uniformity of time periods, and heterogeneity of clinical records have been under-represented in the existing methods. These shortcomings provide reasons why a unified framework that takes effective account of temporal dynamics and relational dependencies is necessary.

Against this backdrop, this paper has suggested a Graph Neural Networks-based model that transforms longitudinal clinical data into dynamic graphs, which can be fully trained to learn the patterns of disease progression.

3. METHODOLOGY

3.1 Problem Formulation

Represent the longitudinal clinical record of a patient as a multivariate time series:

$$X = \{x_1, x_2, \dots, x_T\}, x_t \in \mathbb{R}^d \quad (1)$$

where x_t denotes the clinical feature vector at time step t, and d represents the number of clinical variables, including laboratory measurements, vital signs, and diagnostic indicators. The sequence itself is not necessarily evenly spaced in time because of irregular sampling, and missing observations.

The disease progression modeling objective is to learn a mapping function:

$$f: X \to y \quad (2)$$

where y denotes the disease progression outcome, which may correspond to discrete stage classification or continuous risk estimation. The aim is to learn this role in the sense that the temporal dependencies as well as inter features dependencies are maintained.

3.2 Graph Construction

To capture the complex dependencies among clinical variables, the longitudinal data are transformed into a graph structure $G = (V, E)$, where each node $v_i \in V$ corresponds to a clinical feature. Edges $e_{ij} \in E$ represent relationships between features, which may include statistical correlations, physiological interactions, or temporal co-occurrence patterns.Fig. 1 shows that raw longitudinal clinical data are initially converted to a graph representation at each time

step, with nodes representing clinical variables, and edges representing their dependencies.

The adjacency matrix $A \in \mathbb{R}^{d \times d}$ defines the graph connectivity such that:

$$A_{ij} = \begin{cases} 1, & \text{if a relationship exists between nodes } i \text{ and } j \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Pearson correlation coefficient between clinical features is used to construct edges in this study. Specifically, an edge between nodes i and j is established if the absolute correlation value exceeds a predefined threshold $\tau = 0.6$, i.e., $| r_{ij} | > \tau$.

The latter makes sure that only statistically significant relationships are retained to eliminate noise and enhance the quality of the graph representation.

To incorporate temporal dynamics, a sequence of graphs $\{G_1, G_2, \ldots, G_T\}$ is constructed, where each graph corresponds to a time step t and shares the same topology while allowing node features to evolve over time.

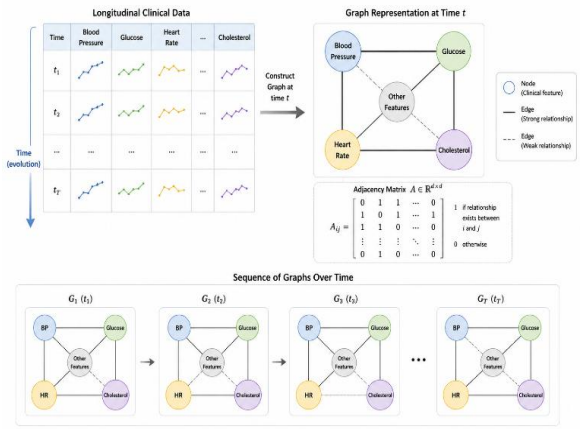


Fig. 1. Graph Construction from Longitudinal Clinical Data for Disease Progression Modeling

3.3 Graph Neural Network Model

For each time step t, the graph G_t is processed using a Graph Neural Network to learn node-level embeddings. The propagation rule for the graph convolution operation is given by:

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \quad (4)$$

where $H^{(l)} \in \mathbb{R}^{d \times k}$ denotes the node feature matrix at layer l, $W^{(l)}$ represents trainable weight parameters, D is the degree matrix of A, and $\sigma(\cdot)$ is a nonlinear activation function such as ReLU.

This formulation will guarantee every node updates its representation with information aggregated by its neighboring nodes, thus capturing inter-feature dependencies.The output of the final GNN layer at time step t produces a latent representation $\square_t$, which encodes the structural relationships among clinical features.

3.4 Temporal Aggregation Mechanism

To model disease progression across time, the sequence of graph embeddings $\{\square_1, \square_2, \ldots, \square_T\}$ is aggregated into a unified patient representation. Fig. 2 shows that a graph embedding is generated with each time step, and is then weighted by its clinical significance and aggregated. A time aggregation scheme is used which is weighted:

$$\square = \sum_{t=1}^{T} \alpha_t \, \square_t \quad (5)$$

where α_t denotes the importance weight assigned to the embedding at time step t. These weights are learnt in the course of training whereby the model is able to concentrate on time points that are clinically relevant. This mechanism helps the framework to get the short term swings and long term trend of progression of patient data.

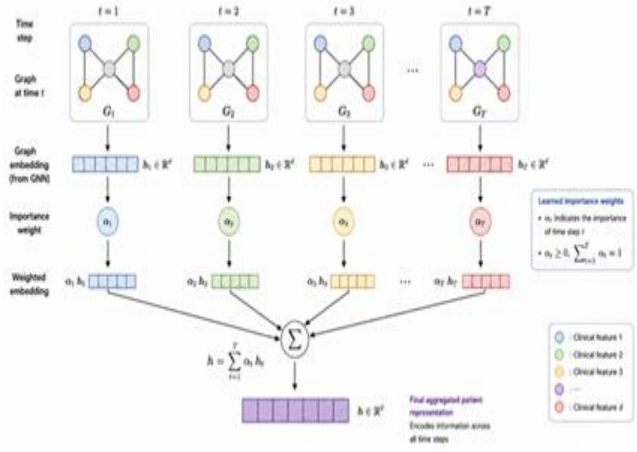


Fig. 2. Temporal Aggregation Mechanism for Disease Progression Modeling Using Graph Embeddings

3.5 Prediction Layer and Loss Function

The aggregated representation $\square$ is passed through a fully connected layer followed by a softmax activation to obtain the predicted output $\hat{y}$. For classification tasks, the model is trained using the cross-entropy loss function:

$$\square = -\sum_{i=1}^{N} y_i \log(\square \hat{y}_i) \quad (6)$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability for class i. This goal objective promotes the model to give predictions that are close to the actual disease progression labels.

3.6 Training Procedure

Backpropagation of stochastic gradient descent is used to train the model. Parameters are optimized to optimize the loss function on training data. The methods of regularization, dropout and weight

decay, are used to avert overfitting. The training process will be maintained until convergence on the basis of the validation performance is reached.

4. Experimental Setup

4.1 Dataset

The presented framework will be tested on the publicly available MIMIC-III (Medical Information Mart for Intensive Care III) clinical dataset that comprises of the detailed longitudinal health records of patients in intensive care unit (ICU) units. A random sample of some 5,000 patients is tracked down using data availability and completeness. The patient records are multivariate time-series data containing vital signs, laboratory test results, and diagnosis codes, or 10 to 50 time steps. Because of the irregularity of the clinical data and the absence of the values, preprocessing procedures are utilized, such as data imputation, data normalization, time-tempor feature alignment. A 70:15:15 split of the dataset is used to create training, validation, and testing sets with mutually exclusive patient records across splits. This segregation helps avoid data leakage and provide credible performance measurement.

4.2 Baseline Models

To confirm the usefulness of the developed Graph Neural Network model, they are compared with a number of baseline approaches. The next step involves establishing the baseline performance of linear and nonlinear classification tasks with the help of the traditional machine learning models, such as Logistic Regression and Random Forest. Furthermore, sequence models based on deep learning, including Long Short-Term Memory (LSTM) and gated recurrent unit (GRU) networks, are used to learn temporal features of longitudinal data. These models are powerful standards of assessing the benefits of graph-based representations. All the models are trained in the same experimental environment in order to provide fair comparison.

4.3 Training Configuration

The proposed GNN model and baseline deep learning models are trained with Adam optimizer with a learning rate of 0.001. It is trained with 100 epochs, where early stopping is used, depending on the loss on validation to avoid overtraining. Efficient gradient updates are made with a batch size of 32. Regularization methods, such as dropout and L2 weight decay are added to enhance the performance of generalization. The choice of hyperparameters is done considering the performance on the validation set to allow the model to be configured to achieve the best performance.

4.4 Evaluation Metrics

The proposed model is measured with a number of standard measures of its performance, such as accuracy, precision, recall, F1-score, and ROC-AUC, to gain a complete understanding of the performance of the suggested model. The generally correctness of predictions is measured by accuracy, and the ability of the model to identify positive instances correctly and detect relevant instances is measured by precision and recall. The F1-score is a balanced measure, which uses precision and recall. Also, ROC-AUC is to determine the discriminative power of the model to different decision thresholds, which is especially relevant in healthcare-related cases when the dataset is not balanced.

4.5 Implementation Details

All simulations are carried out with the PyTorch deep learning platform. The system that the models are trained on has NVIDIA GPU (e.g., Tesla T4 / RTX 3060) and a standard data preprocessing CPU environment. GPU acceleration drastically decreases the time of training and allows to process the graph-based computations effectively.

5. RESULTS AND DISCUSSION

5.1 Quantitative Results

The proposed Graph Neural Network (GNN) model is compared to the baseline methods such as Logistic Regression, Random Forest, LSTM, and GRU. Table 1 summarizes the results.

Table 1. Performance Comparison of Models

Model	Accuracy	F1-score	ROC-AUC
Logistic Regression	0.74	0.72	0.75
Random Forest	0.78	0.76	0.80
LSTM	0.82	0.80	0.84
Proposed GNN	0.87	0.85	0.89

The proposed GNN model also demonstrates the greatest level of performance based on all evaluation metrics, improving its accuracy and ROC-AUC by around 5-9 per cent more than the traditional machine learning models, and 3-5 per cent more than the sequence-based deep learning models like LSTM. Fig. 3 shows the comparative performance using various evaluation metrics, thus obviously proving the superiority of the proposed GNN model over the approaches used as baseline. Moreover, Receiver Operating Characteristic (ROC) analysis is used as a measure of performance in classification. As demonstrated in Fig. 4, the proposed GNN has greatest Area Under the Curve (AUC), which means that it offers an improved trade-off between sensitivity and specificity than the baseline models. This confirms

the high level of discrimination in the model in the tasks of predicting the progress of the disease.

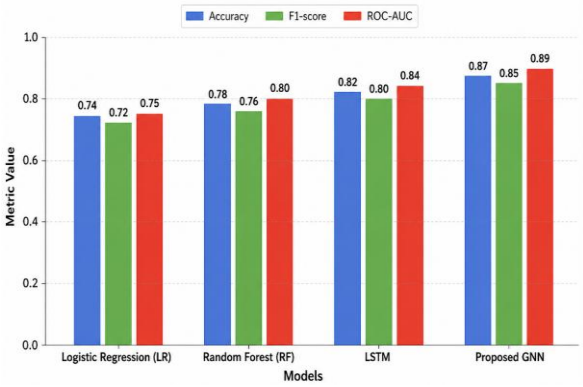


Fig. 3. Performance Comparison of Models Across Evaluation Metrics

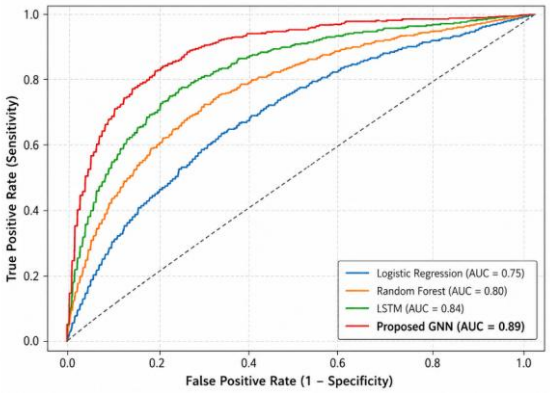


Fig. 4. ROC Curve Comparison of Proposed GNN and Baseline Models

5.2 Comparative Analysis

High performance of the proposed GNN model can be explained by the fact that this model can simultaneously learn relational relationships and temporal dynamics that is vital in longitudinal clinical data. Logistic Regression and other traditional models do not deal with nonlinear interactions of clinical variables, which leads to reduced predictive performance. Random Forest is better as it models nonlinear relationships and still fails to address the issue of modeling temporal dependencies. The models that are based on sequences like LSTM perform better because they are able to model the trends of time. Nevertheless, they do not explicitly learn inter- features relationships and handle input features separately. Conversely, the proposed GNN framework uses graph-based representations to encode interaction of features such that the model can pick up more informative embeddings. Moreover, the GNN model is suitable in irregular sampling over time as it combines the information of the graph at various time steps, which makes it strong in real-life clinical application. These results can be compared to the recent research in healthcare machine learning, which emphasizes the significance of structural and temporal modeling to achieve better predictive accuracy.

5.3 Statistical and Clinical Interpretation

Statistically, the improvements in ROC-AUC that were observed show that the proposed model is more capable of discriminating between different disease progression conditions across different decision thresholds. This is especially essential in clinical practice where there is a need to balance sensitivity and specificity. The increase in the F1-score also indicates that the model can have a more optimal trade-off between the precision and the recall and minimize the number of false positives and false negativities. This is important in modeling disease progression, in which poor predictions can result in missed treatment or unwarranted intervention.

5.4 Interpretability Analysis

The major benefit of the suggested framework is that it is interpretable. The model is able to detect clinically significant variables, which cause disease progression, by looking at graph-based features interaction and attention weights. As an illustration, vital signs and laboratory indicators that have a high degree of connectivity in the graph structure are likely to have greater impact on the results of the prediction. This gives useful data to clinicians and helps them gain a clearer view of the dynamics of diseases and facilitates making decisions based on data.

5.5 Discussion

These findings indicate that graph-based learning combined with a temporal modeling approach improves predictive performance in longitudinal healthcare data greatly. The suggested GNN model exhibits a better accuracy, as well as better interpretability than an existing method. But the quality of the graph construction and feature selection determine the performance of the model. Untrue edgeformations or noisy data can impact on the learning efficiency. Also, it has a higher computational complexity than traditional models, which can affect very large datasets.

CONCLUSION AND FUTURE WORK

The current paper has introduced a Graph Neural Network (GNN)-based model to predict disease progression, using longitudinal clinical data. The proposed method successfully captures inter-feature and temporal dependencies, unlike traditional and sequence-based models, by turning time-series clinical records into graph representations that are dynamic. The combination of both the graph convolution and temporal aggregation mechanism allows the model to come

up with meaningful representations of how diseases change with time. Exposure to the MIMIC-III data shows that the proposed framework is able to achieve better results in comparison to baseline models, such as Logistic Regression, Random Forest, and LSTM by a margin of up to 5–9 percentage points in accuracy and ROC-AUC. These results affirm the usefulness of graph-based learning in processing heterogeneous, irregular and high-dimensional clinical data. Besides the superior predictive accuracy, the model offers a better interpretability of the clinically significant interaction of features, which gives it relevance to the real-world healthcare decision-making. Although such good outcomes are achieved, there are nonetheless several challenges. The existing framework is based on the pre-defined graph structures, which might restrict flexibility in the extremely dynamic clinical setting. Moreover, it should be scalable and computationally complex to be used on a large scale. The further development of the presented model will be aimed at its expansion in a number of directions. First, additional modeling of the long-range dependencies can be based on incorporating Transformer-based temporal mechanisms. Second, the use of multimodal data sources, including medical imaging and genomic data, can enhance the prediction performance and clinical significance. Third, attempts will be done on real-time implementation in the clinical decision support systems so that real implementation of the technology can be adopted in the healthcare environment. Lastly, the framework can also be applied to model multi-disease progression, where one can concomitantly analyze a number of comorbid conditions.

REFERENCES

1. C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik, "Evaluating explainability for graph neural networks," *Scientific Data*, vol. 10, no. 1, p. 144, 2023.
2. X. Dong *et al.*, "An integrated LSTM-HeteroRGNN model for interpretable opioid overdose risk prediction," *Artificial Intelligence in Medicine*, vol. 135, p. 102439, 2023.
3. Y. Li, B. Qian, X. Zhang, and H. Liu, "Graph neural network-based diagnosis prediction," *Big Data*, vol. 8, no. 5, pp. 379–390, 2020.
4. Y. Li, B. Qian, X. Zhang, and H. Liu, "Knowledge guided diagnosis prediction via graph spatial-temporal network," in *Proc. SIAM SDM*, 2020, pp. 19–27.
5. B. T. Lee, O. Y. Kwon, H. Park, K. J. Cho, J. M. Kwon, and Y. Lee, "Graph convolutional networks-based noisy data imputation in electronic health record," *Critical Care Medicine*, vol. 48, no. 11, pp. e1106–e1111, 2020.
6. D. Lee, X. Jiang, and H. Yu, "Harmonized representation learning on dynamic EHR graphs," *Journal of Biomedical Informatics*, vol. 106, p. 103426, 2020.
7. Z. Sun, H. Yin, H. Chen, T. Chen, L. Cui, and F. Yang, "Disease prediction via graph neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 818–826, 2021.
8. Y. Si *et al.*, "Deep representation learning of patient data from electronic health records (EHR): A systematic review," *Journal of Biomedical Informatics*, vol. 115, p. 103671, 2021.
9. H. N. Cho *et al.*, "Heterogeneous graph construction and HinSAGE learning from electronic medical records," *Scientific Reports*, vol. 12, no. 1, p. 21152, 2022.
10. S. Tang *et al.*, "Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks," *IEEE Journal of Biomedical and Health Informatics*, 2023.
11. X. Wang and J. Liu, "TAGNet: Temporal aware graph convolution network for clinical information extraction," in *Proc. IEEE BIBM*, 2020, pp. 2105–2108.
12. Y. Xu *et al.*, "Time-aware context-gated graph attention network for clinical risk prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022.
13. F. Xie *et al.*, "Deep learning for temporal data representation in electronic health records: A systematic review," *Journal of Biomedical Informatics*, vol. 126, p. 103980, 2022.
14. Y. Xu, H. Ying, S. Qian, F. Zhuang, X. Zhang, D. Wang, J. Wu, and H. Xiong, "Time-aware context-gated graph attention network for clinical risk prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022.
15. H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," 2022.
16. I. A. Chikwendu *et al.*, "A comprehensive survey on deep graph representation learning methods," *Journal of Artificial Intelligence Research*, vol. 78, pp. 287–356, 2023.
17. H. Lu and S. Uddin, "Disease prediction using graph machine learning based on electronic health data: A review of approaches and trends," *Healthcare*, vol. 11, no. 7, 2023.