Computers in Biology and Medicine 183 (2024) 109282

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Graph neural network-based subgraph analysis for predicting adverse drug events

Fangyu Zhou [a], Matloob Khushi [b,c], Jonathan Brett [d,e], Shahadat Uddin [a,*]

[a] School of Project Management, Faculty of Engineering, The University of Sydney, Australia
[b] School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
[c] Department of Computer Science, Brunel University London, Uxbridge, London, UK
[d] St Vincent's Clinical School, The University of New South Wales, Sydney, New South Wales, Australia
[e] Department of Clinical Pharmacology and Toxicology, St Vincent's Hospital Sydney, Sydney, New South Wales, Australia

ARTICLE INFO

Keywords:
Adverse drug events
Graph neural network
Machine learning
Administrative data

ABSTRACT

Purpose: Adverse drug events (ADEs) are a significant global public health concern, and they have resulted in high rates of hospital admissions, morbidity, and mortality. Prior to the use of machine learning and deep learning methods, ADEs may not become well recognized until long after a drug has been approved and is widely used, which poses a significant challenge for ensuring patient safety. Consequently, there is a need to develop computational approaches for earlier identification of ADEs not detected during pre-registration clinical trials.
Methods: This paper presents a state-of-the-art network-based approach that models patients as subgraphs composed of nodes of International Classification of Diseases (ICD) codes and directed edges illustrating disease progression. Four Graph Neural Network (GNN) variants were employed to make sub-graph level predictions that answer three Research Questions (RQ): 1) whether ADE(s) would occur given a patient's prior diagnoses history, 2) when an ADE would occur, and 3) which ADE would occur. The first and second RQs were addressed using a binary classification approach. The third RQ was addressed using a multi-label classification model.
Results: The proposed network-based approach demonstrated superior performance in predicting ADEs, with the GraphSage model exhibiting the highest accuracy for both RQ 1 (0.8863) and RQ 3 (0.9367), while the Graph Attention Networks (GAT) model was found to perform best for RQ 2 (0.8769). Furthermore, an analysis segmented by ADE classification revealed that while RQs 1 and 3 exhibited minimal variance across different ADE categories, a distinct advantage was observed for categories B, C, and E in the context of RQ 2 when applying this sub-graph method.
Conclusion: The network-based approach demonstrates the potential of GNNs in supporting the early detection and prevention of ADEs. Accurately predicting ADEs could enable healthcare professionals to make informed clinical decisions, take preventive measures and adjust medication regimens before serious adverse events occur. The proposed prediction method could also lead to optimized usage of healthcare resources by preventing hospital admissions and reducing the overall burden of adverse drug events on the healthcare systems.

1. Introduction

The issue of medication safety has become an increasingly pressing concern in many countries. A multitude of studies conducted across various regions have revealed that adverse drug reactions (ADRs) and/or adverse drug events (ADEs) are closely associated with a significant proportion of hospital admissions, ranging from 3.7 % to 16.6 % [1]. Hospital readmission rates resulting from drug-related incidents have been documented to vary between 3 % and 64 %, with a median of 21 % [2]. In Australia, where a substantial proportion of the population (47 %) is reported to consume prescription drugs each fortnight, adverse drug events result in hospitalization for 1.3 %–4.6 % of patients and up to 9 % of emergency admissions [3]. Additionally, it has been reported that adverse drug reactions result in the death of approximately 197,000 people in Europe each year [4]. These issues have also resulted in significant direct and indirect economic burdens for numerous countries

* Corresponding author. 21 Ross Street, Forest Lodge, Sydney, NSW, 2037, Australia.
E-mail addresses: fangyu.zhou@sydney.edu.au (F. Zhou), matloob.khushi@brunel.ac.uk (M. Khushi), j.brett@unsw.edu.au (J. Brett), shahadat.uddin@sydney.edu.au (S. Uddin).

https://doi.org/10.1016/j.compbiomed.2024.109282
Received 24 May 2024; Received in revised form 2 October 2024; Accepted 14 October 2024
Available online 23 October 2024

[5].

The World Health Organization (WHO) defines ADRs as noxious and unintended responses that occur at doses generally used in the human body [6]. Adverse drug events (ADEs) are defined more broadly as any injuries caused by medications, including overdoses and medication errors [7]. Distinguishing between ADRs and ADEs can be clinically challenging due to insufficient information available, and this follows through to how such hospital encounters are coded following discharge. Therefore, this study will not differentiate between these two terms but will refer to them collectively as ADEs that are identified using ICD-10 diagnostic codes through a systematic review conducted by Hohl et al. [8]. The details of their classifications will be presented in the Data Preparation section.

The use of machine learning and deep learning methods in predicting ADEs has the potential to revolutionize clinical practice by providing early detection and prevention of adverse drug reactions [9]. For instance, Pauwels et al. [10] investigated the chemical substructures of drugs and developed prediction models at the pre-clinical stage using k-nearest neighbors (kNN), support vector machine (SVM), ordinary canonical correlation analysis (OCCA), and sparse canonical correlation analysis (SCCA). Similarly, Liu et al. [11] employed five types of drug feature vectors and applied logistic regression (LR), naïve Bayes (NB), kNN, random forest (RF), and SVM to build their prediction models. Huang et al. [12] considered drug targets, protein-protein interaction networks, and gene ontology annotation and employed two classifiers - SVM and LR.

In recent studies, many researchers used graph-based methods to identify ADRs and ADEs from different data sources. GNN and its variants have shown outstanding applicability and scalability when tailored for various prediction tasks [13]. Notably, Bean et al. [14] proposed a comprehensive knowledge graph of four types of nodes: drugs, protein targets, indications, and adverse reactions. Enrichment tests were applied to this graph to learn the characteristics of drugs and predict ADRs. Zitnik et al. [15] developed a novel poly-medication graph to predict ADRs resulting from given drug combinations. They utilized convolutional neural network methods to explore latent information between drugs and proteins, drugs to drugs, and proteins to proteins. In another study, Deac et al. [16] proposed a graph neural network (GNN) based on the common attention mechanism. ADR types and drug molecular structures were used to predict possible ADEs resulting from multidrug combinations. Yu [17] developed a deep multi-structured neural network model based on multi-scaled features by utilizing sequence-based word embedding, substructure-based molecular fingerprint, and chemical structure-based graph embeddings. Recently, more GNN-based models have emerged. Chen [18] introduced an innovative Graph Neural Network architecture. This sophisticated approach integrates two key components: one based on Drug Chemical Structure Graphs and another leveraging Drug Knowledge Graphs. By synthesizing these elements, the model effectively captures the multimodal characteristics of pharmaceutical compounds for a better drug to drug predictions. Khan et al. [19] proposed a new method of generating polymedication regimen and polymedication networks from the administrative data, which could be used as the basis for further detection of adverse drug reactions. It was proven in a study by Wang [20] that the involvement of sub-graph can help improve the ADE predictions as this method provides more explainable paths. A recent comprehensive review of the datasets and approaches used to identify ADRs [21] suggests that researchers have made significant progress by leveraging comprehensive information from molecular drug structures, drug attributes such as drug pathways, protein interactions, and other data sources. A summary of these related works could be found in Table 1. While all these studies have made significant progress in predicting adverse drug reactions using various machine learning techniques and data sources, there remains a gap in utilizing patient diagnosis history for ADE prediction, which comes as a research motivation for us to aim to fill this gap by proposing a novel graph-based modeling method that uses health claims data and patients' ICD (International Classification of Diseases) code histories.

Zhou and Uddin proposed a subgraph prediction method for ADR that uses information from connected patients' diagnosis histories and the topological structure of a patient's history [22]. The current study deepens this research by varying the research questions and endeavors to solve three related problems. This approach represents a paradigm shift in ADE prediction by viewing each patient's medical history as a subgraph within a larger graph of disease progression. The contributions of this research are threefold. First, this is the first study that uses only patients' diagnosis history to predict ADEs by modeling each patient's history as a subgraph that lies in a full graph where ICD codes are used as nodes, and the directed edges illustrate the progression of patients' diseases. Second, administrative data is exclusively used as an enabling source for early prediction of ADEs, which is time-efficient and cost-effective in comparison to other computationally expensive methods. As has been pointed out in a recent review paper by Luo [23], extensive and rich drug features can undoubtedly improve model performance. Still, they could also introduce more noise and, therefore, a dataset where more readily accessible negative samples emerge as a crucial factor in enhancing predictions of ADEs. This gap shall be moderately bridged by the proposed dataset used in this study. This choice of data also distinguishes our work from most existing studies in the field, which typically incorporate various drug-related datasets. Third, the proposed framework not only predicts whether ADEs are associated with a specific patient but also identifies when and which ADEs occur with high accuracy.

This paper aims at three research questions.

Research Question 1: Can graph machine learning methods distinguish between patients who experience ADEs and patients who do not?
Research Question 2: During a particular admission, can this method be used to predict the likelihood of ADEs based on their prior diagnosis?
Research Question 3: What type(s) of ADEs are likely to develop?

RQ 1 helps us understand how well graph machine learning methods can predict ADEs. RQ 2 assists hospitals or clinics in predicting the likelihood of a patient experiencing an ADE during a particular admission or clinical encounter. RQ 3 aids in understanding which specific ADEs are most likely to occur in patients.

The subsequent sections are structured as follows: The 'Data Preparation' section outlines the data source, cohort selection, and data labeling methods. The 'Methods' section describes the GNN-based methods employed to predict ADEs. The 'Results' section empirically evaluates the proposed framework's efficacy for the three ADE prediction questions. Following this, we discuss the findings and limitations of the experiments and highlight potential avenues for future research. Finally, we conclude the study in the 'Conclusion' section.

2. Data preparation

This study draws upon health claims data from the Commonwealth Bank Health Society (CBHS), an Australian private health insurance company [24], to analyze patients' progression of diseases over time using ICD codes. A flowchart that details the selection process and the corresponding number of patients at each stage is shown in Fig. 1. The dataset includes 419,952 de-identified patients from 1976 to 2018, of which 123,983 have at least one claim record. The dataset records patients' ID, age and gender and also each claim's details, including service type (ICD-9 or ICD-10), item number (ICD codes), created date, and diagnosis procedure code, among others. This study focuses only on patients recorded with ICD-10 codes to ensure consistency, and 118,695 patients were identified to have ICD-10 codes recorded following hospital admissions. The ICD-10 codes were utilized as inputs for the Graph

Table 1
Summary of related papers.

Ref	Title	Dataset	Method	Evaluation metric	Pros	Cons
[10]	Predicting drug side-effect profiles: a chemical fragment-based approach	SIDER, DrugBank	Sparse canonical correlation analysis	Accuracy, Area Under the ROC Curve (AUROC)	The unified framework integrates chemical and pharmacological spaces. This integration enables the extraction of correlated sets of chemical substructures and side effects. It allows for the simultaneous prediction of numerous potential side effects.	These processes rely heavily on predefined chemical substructures and specific side-effect terminology.
[11]	Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs	SIDER, DrugBank, PubChem, KEGG	LR, NB, kNN, RF, SVM	Accuracy, Precision, Recall, Area Under the ROC Curve (AUROC)	Chemical properties (compound signatures), biological elements (targets, transporters, enzymes, and pathways), and phenotypic characteristics (indications and known side effects) are integrated into this approach.	The phenotypic features were represented in a relatively simple manner. More complex techniques, such as categorizing drug indications through ontologies, warrant further investigation. Moreover, a drug's action involves perturbing biological systems, encompassing various molecular interactions. These interactions include protein-protein interactions, signaling pathways, and pathways related to drug action and metabolism.
[12]	Predicting adverse side effects of drugs	SIDER, DrugBank, HAPPI	Integrating gene network and gene annotation	Accuracy, Precision, Recall, Area Under the ROC Curve (AUROC)	Clinical observation data is combined with drug target information, protein-protein interaction (PPI) networks, and gene ontology (GO) annotations.	This approach enables the examination of functional relationships between proteins lacking direct associations. Furthermore, experimentally derived genotype-phenotype data, such as that obtained from genome-wide association studies, may prove valuable. Recent studies on genetic polymorphisms of cardiotoxicity-inducing enzymes have already demonstrated the potential of this additional information.
[14]	Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records	SIDER, DrugBank, PubChem, Te EHR at the South London and Maudsley NHS Foundation Trust	Youden's J statistic optimized logistic regression	Area Under the ROC Curve (AUROC)	A knowledge graph was constructed, comprising four node types: drugs, protein targets, indications, and adverse reactions. Utilizing this graph, they developed a machine learning algorithm founded on a basic enrichment test. Initial demonstrations showed this method's exceptional performance in classifying known causes of adverse reactions.	A significant constraint of this approach lies in the necessary assumption that patients adhere to their prescribed medications. Additionally, they must assume that all medications taken by patients are accurately captured in the Electronic Health Record (EHR).
[15]	Modeling polypharmacy side effects with graph convolutional networks	STITCH, SIDER, OFFSIDES, TWO SIDES	Decagon model	Area Under the ROC Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC)	Decagon's functionality includes predicting associations between side effects and co-prescribed drug pairs (drug combinations). This capability allows for the identification of side effects that are not attributable to individual drugs in isolation.	The methodology incorporates molecular protein-protein and drug-target networks in conjunction with population-level side effect data from patients. Additional biomedical information sources, such as dosed drug concentration levels, may be pertinent to modeling the side effects of drug pairs. We anticipate exploring the potential benefits of integrating these additional data sources into the model.
[16]	Drug-drug adverse effect prediction with graph co-attention	STITCH, SIDER, OFFSIDES, TWO SIDES	MHCADDI model	Area Under the ROC Curve (AUROC)	A graph neural network architecture has been introduced, achieving state-of-the-art results in predicting potential polypharmacy side effects from drug combinations. This architecture relies exclusively on the molecular structure information of drug pairs.	Potential future research could explore the application of similar cross-modal architectures to predict interactions in diverse network types. These could include language or social networks, where components from different networks interact with one another.

(*continued on next page*)

Table 1 (*continued*)

Ref	Title	Dataset	Method	Evaluation metric	Pros	Cons
[17]	MSDSE: Predicting drug-side effects based on multi-scale features and deep multi-structure neural network	SIDER, Drugbank, PubChem, MedDRA	MSDSE model	Area Under the Precision-Recall Curve (AUPRC), mean reciprocal rank (MRR), F1, Matthews correlation coefficient (MCC)	Diverse features are fused to provide a more comprehensive portrayal of drug properties. Additionally, an adaptive neural network is designed to align with the feature data structure for enhanced processing. The resulting high-quality features facilitate subsequent prediction tasks.	A significant constraint of these approaches is their applicability only to drugs in the maturation stage.
[18]	An effective framework for predicting drug–drug interactions based on molecular substructures and knowledge graph neural network	Drugbank, KEGG	MSKG-DDI model	Accuracy, F1, Area Under the Precision-Recall Curve (AUPRC), Area Under the ROC Curve (AUROC)	MSKG-DDI uses knowledge-embedded neural networks on raw molecular graphs to extract rich drug features. It predicts drug-drug interactions by identifying substructure interactions, enabling predictions for new drug pairs.	Future work could focus on improving model interpretability through visualization and explanation methods and identifying key components in molecular and knowledge graphs for DDI predictions.
[20]	Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning	Drugbank, TWO SIDES	KnowDDI	F1, Accuracy, Cohen's k	They aim to develop an effective Drug-Drug Interation (DDI) predictor from rare DDI fact triplets. KnowDDI leverages biomedical knowledge and deep learning to enhance drug representations and similarities, compensating for the lack of known DDIs.	They exclude molecular features of drugs to test KnowDDI's ability to learn from external knowledge graphs and DDI fact triplets alone.

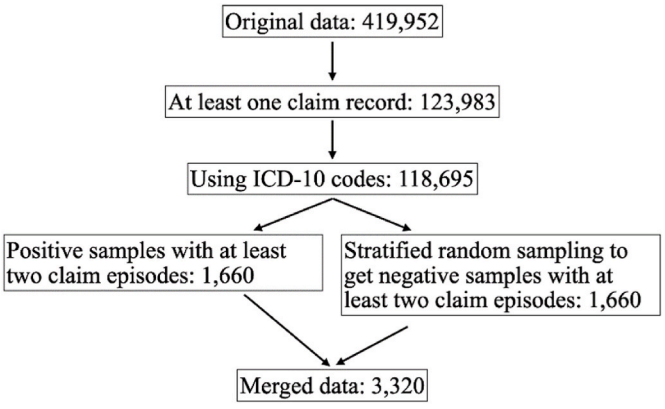

Fig. 1. Study cohort selection process.

Neural Network algorithm by modeling each patient as a subgraph. In this model, each node represents an ICD-10 code, and each edge signifies the progression of diagnoses over time. Detailed modeling methods and a simplified example are in Section 3.1: ICD-10 Code Network.

Further filtering includes that only individuals who had a minimum of two claim episodes were considered, and patients were filtered based on the ADEs codes identified in the previously mentioned systematic review [8]. After these filtering steps, 1660 patients who were labeled with at least one ADEs code were identified. For a fair comparison, selecting 1660 patients from the eligible individuals who did not experience any ADEs was challenging since it is often difficult to find patients with identical demographic information, but we attempted to overcome this challenge by selecting patients based on similar age and gender distributions. Specifically, we stratified our cohorts into distinct segments based on gender and age intervals of five years. To maintain a balanced representation, we employed a stratified sampling method to identify the corresponding number of negative patient cases within each segment. In this way, we ensured that our classification captured the unique characteristics of each age and gender group and that our analysis accounted for potential differences between these groups. Table 2 displays detailed statistics regarding the age distribution of the two genders.

Table 2
Age distribution of the studied cohort (3320 patients), categorized by gender.

	Count	Average	Maximum	Minimum	Standard Deviation
Male	1450	51.96	106	1	25.80
Female	1870	50.22	106	0	22.80

As pointed out by several scholars, establishing a causal link between an ADE and drug exposure can be a daunting task due to the existence of other potential risk factors such as medical care, underlying conditions, and genetic predispositions [25]. Notably, no universally accepted framework for assessing ADE causality is present [26]. There are tools used clinically, e.g., Naranjo score, but causal inference at a population level requires several considerations. Nonetheless, the study by Hohl et al. [8] contains a causality rating table in the appendix section that features nine categories, with A1, A2, B1, and B2 being four of those categories that directly relate to medication or drug usage. Categories C (very likely), D (likely), E (possible), and U (unlikely) signify varying degrees of the likelihood that the ICD-10 code corresponds with an ADE, while category V represents Vaccine-associated. Importantly, only categories A1, A2, B1, B2, and C, D, and E are regarded as validated to have a causative link, which was used in this study, accordingly. These segments, labeled A, B, C, D, and E, were defined based on distinct criteria outlined in Table 3.

To answer the three RQs posed earlier, the selected cohorts were labeled differently. A simplified example is shown in Table 4, which displays the ICD code (represented by random letters from 'A' to 'N') history of four patients, with each patient corresponding to a row and the time sequence represented from left to right. The bold and underlined letters 'F' and 'G' denote ADEs. This table also addresses the three distinct RQs as abovementioned, and are labeled accordingly.

In RQ 1, the labels correspond to which group of cohorts the patients are in. The RQ 2 aims to predict when a patient will experience ADEs, given that they have been classified as ADE-associated patients based on the criteria established in RQ 1. The patient histories for Patient ii and Patient iii are divided into (n-1) segments, where n is the total number of their ICD codes. If the last ICD-10 code belongs to a class of ADE, this will be labeled as 1; otherwise, it will be labeled as 0. In RQ 3, an ADE will be

Table 3
Classification of ICD-10 codes of ADEs with examples.

Code Category	Definition	Code example	Code example description	Number of code counts in the entire cohort	Number of distinct patient counts
A	The ICD-10 code description includes the phrase 'induced by medication/drug' or 'induced by medication or other causes'.	J70.2 142.7	Acute drug-induced interstitial lung disorders Cardiomyopathy due to drugs and other external agents	2141	1339
B	The ICD-10 code description includes the phrase 'poisoning by medication' or 'poisoning by or harmful use of medication or other causes'.	T36 X44	Poisoning by systemic antibiotics Accidental poisoning by, and exposure to, other and unspecified drugs, medicaments and biological substances	364	168
C	Adverse drug events are deemed to be very likely, although the ICD-10 code description does not refer to a drug.	L51.2	Toxic epidermal necrolysis	237	209
D	Adverse drug events are deemed to be likely, although the ICD-10 code description does not refer to a drug	N17	Acute renal failure with tubular necrosis	1538	534
E	Adverse drug events are deemed to be possible, although the ICD-10 code dictionary does not refer to a drug	K25	Gastric ulcer	3135	937

Table 4
An example of four patients' ICD-10 Code history, where there might be ADEs represented by the underlined letters (e.g., F and G), and their labeling methods for the three questions. In the label column, '1' denotes the presence of ADEs identified in the above patients' history, while '0' indicates the absence of such events.

Example of four patients' ICD-10 Code history					
Patient	ICD-10 code history				
i	A	B	C	D	
ii	E	***F***	***G***	C	H
iii	I	J	***F***	E	
iv	K	L	M	N	

Labels for RQ 1	
Patient	Label
i	0
ii	1
iii	1
iv	0

Labels for RQ 2	
Patient history segment	Label
ii – 1 (E, F)	1
ii – 2 (E, F, G)	1
ii – 3 (E, F, G, C)	0
ii – 4 (E, F, G, C, H)	0
iii – 1 (I, J)	0
iii – 2 (I, J, F)	1
iii – 3 (I, J, F, E)	0

Labels for RQ 3	
Patient	Label
ii	0, 0, 0, 1, 1, 0, 0, 0, 0, 0
iii	0, 0, 0, 1, 0, 0, 0, 0, 0, 0

precisely identified using one-hot vectors, which will be assigned a length equivalent to the total number of ADEs contained in the dataset.

3. Methods

The baseline method is a consolidated and customized approach from the literature where multiple authors have used centrality measures for graph prediction. Khan et al. [27] used a comorbidity network to understand chronic disease progression. Building on this study, Khan et al. [28] also proposed using several measures extracted from graph theory and social-network theory that look at the prevalence of comorbidities, transition patterns, and clustering to predict type 2 diabetes. Lu et al. [29] represented patients as nodes and their shared diseases as weighted edges to predict diabetes. Additionally, Zhou et al. [30] proposed using multiple graph centrality measures to predict ADRs, which has successfully utilized administrative healthcare data to create a graph where patients are represented as nodes with edges showing their relatedness. Classical machine learning methods such as logistic regression (LR) can be applied to leverage network centrality measures to predict whether an individual would develop ADRs.

In contrast to the above baseline method, the overall framework of the proposed method is illustrated in Fig. 2, and its details will be illustrated in the following sub-sections. Based on this new framework, GNN-based algorithms should be able to learn the relationships between ICD codes and predict a subgraph.

3.1. ICD-10 Code Network

In this network, each sub-ICD code, such as C15.0 for the cervical part of the esophagus, is represented as a node. The progression of a patient's ICD code history is depicted as a subgraph comprising some nodes, and the edges showing the sequence of their codes. For instance, in Fig. 3, four patients are illustrated, mirroring the data provided in Table 4. The ICD codes F and G are identified as ADEs codes, consequently making each patient associated with these two letters as positive. This approach is particularly advantageous because it allows us to utilize information from ICD codes that may not be present in a patient's diagnosis history but are related to codes that are present. Such relationships and implicit information cannot be easily revealed or explored using traditional machine learning algorithms when applied to the benchmarking patient network as shown in Fig. 4. This also gives an intuition that Graph Neural Network will be able to pick up the latent information and make better predictions.

3.2. GNN-based approaches

As demonstrated in the above construction of the proposed network of ICD codes, it is necessary to learn the hidden relationships between ICD codes in order to achieve more accurate predictions. To accomplish this task, GNN-based algorithms are utilized, as they are designed to aggregate information from neighboring nodes. The general mechanism of the GNN approach is displayed in Fig. 5. These GNNs are built by stacking layers, where each layer represents a node that aggregates information from its neighbors that are one extra hop away. The inputs to the GNN at the first layer consist of the input embeddings of the subgraph nodes, represented as vectors of dimension 256. The second layer contains vectors of dimension 128, and the third layer has vectors of dimension 64. Additionally, we employ the top-k pooling method in the GNN. This method selects the top k nodes based on a learned score,

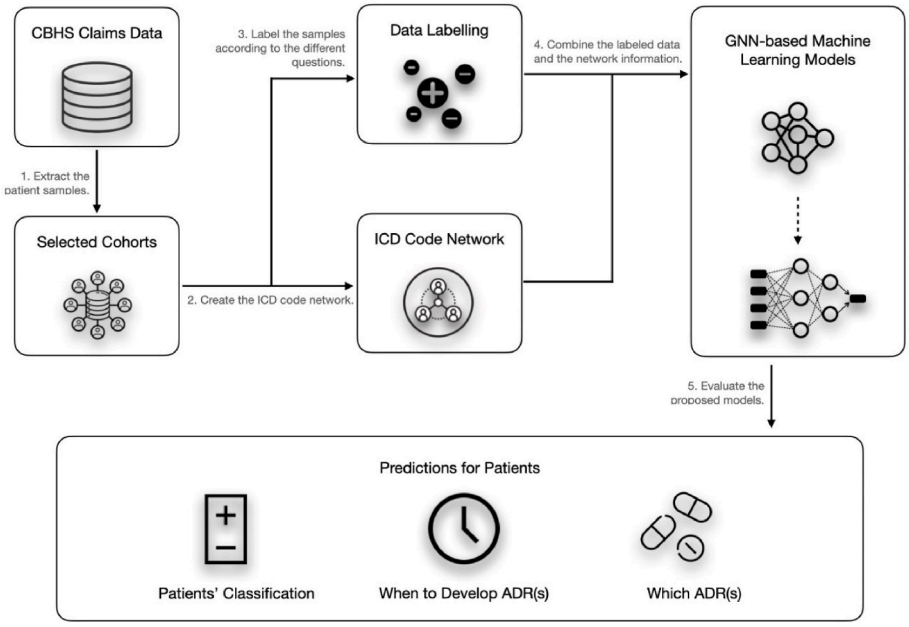

Fig. 2. Overall framework of this study.

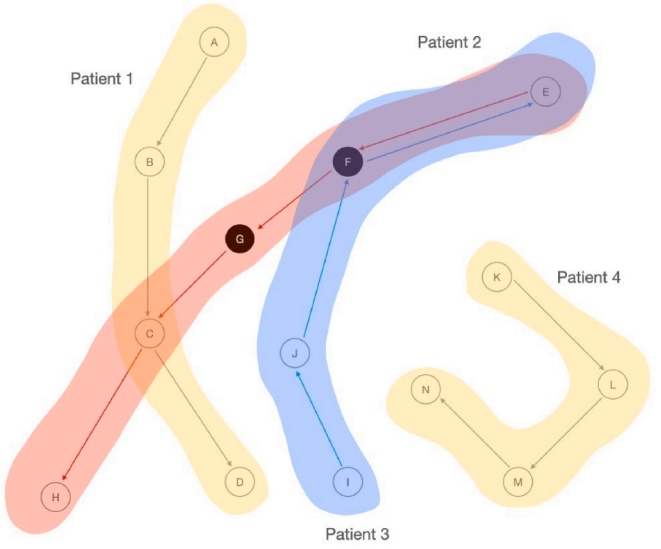

Fig. 3. Proposed ICD-10 codes network, where patients are represented as subgraphs.

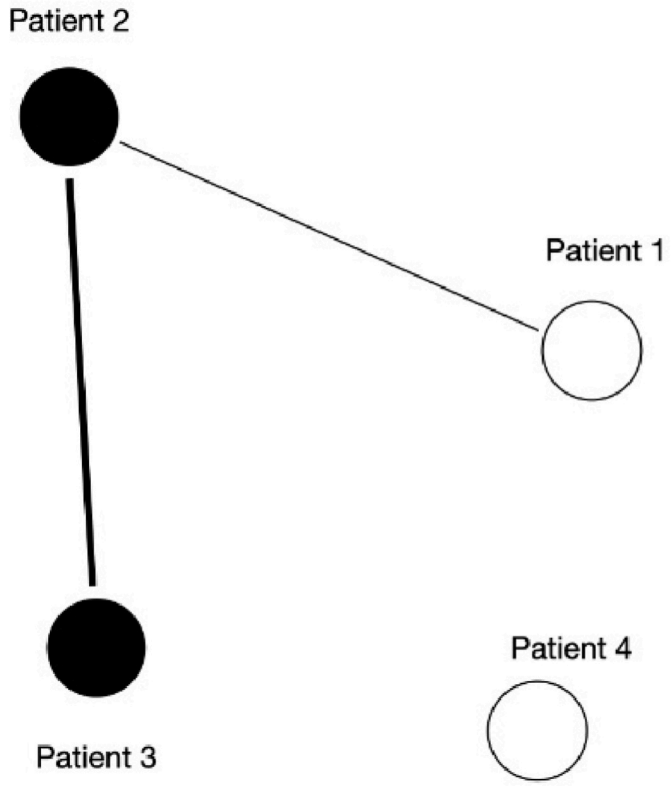

Fig. 4. Benchmark patient network.

allowing the model to focus on the most informative nodes in the graph. By retaining only the most relevant nodes, top-k pooling reduces the complexity of the graph and enhances the model's ability to capture essential patterns and relationships, thereby improving the overall performance of the GNN in tasks such as classification and prediction. Every layer of a GNN embodies three fundamental functions: the message function, aggregation function, and update function.

This study analyses and compares four variants of GNN models, namely: Graph Convolutional Network (GCN) [31], Graph Attention Network (GAT) [32], Graph Attention Network version 2 (GATv2) [33], and Graph Sample and Aggregate (GraphSAGE) [34].

Graph Convolutional Network (GCN) is a fast approximation convolution method used in graph learning. GCN utilizes a layer-wise propagation rule, as shown below [31]:

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

Where $H^{(l+1)}$ refers to the activations of the $(l+1)$ th layer after applying the activation function $\sigma(\cdot)$ (such as ReLU) to the node embeddings in the l th layer; $\widetilde{A}$ represents the adjacency matrix of the constructed

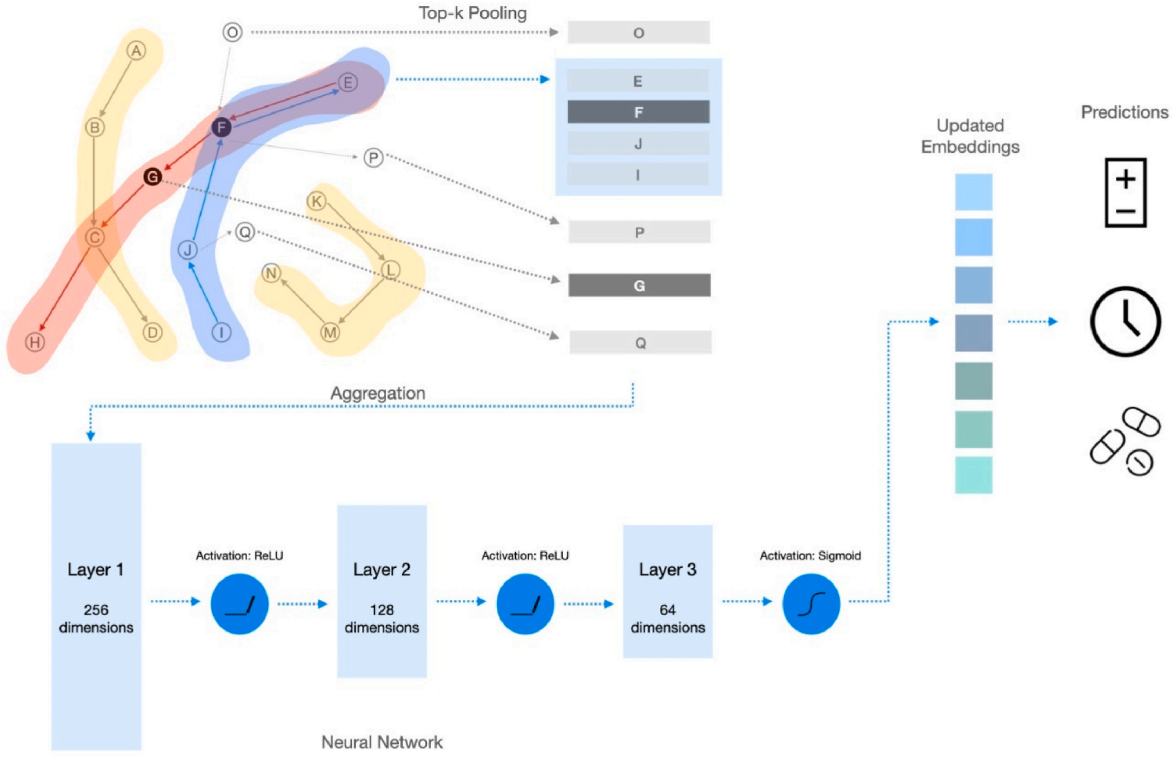

Fig. 5. The general mechanism of GNN-based training methods.

graph, including self-connections for each node (corresponding to the diagonal positions in the matrix); $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and the weight matrix $W^{(l)}$ of the l th layer can be updated during training. $H^{(l)}$ is the activation matrix of the l th layer.

Graph Attention Network (GAT) utilizes an attention mechanism that involves a single-layer feedforward neural network and applies the LeakyReLU nonlinearity [32]. The coefficient for a node pair *(i, j)* can be computed using the formula below:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{a}^T\left[\mathbf{W}\vec{h}_i \,\|\, \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in N_i}\exp\left(\text{LeakyReLU}\left(\vec{a}^T\left[\mathbf{W}\vec{h}_i \,\|\, \mathbf{W}\vec{h}_k\right]\right)\right)}$$

Where T is transpose and $\|$ is vector concatenation; N_i is the first-order neighboring nodes of the node i of the graph; $\vec{h}$ is the input node features; $\mathbf{W}$ is the weight matrix that serves as a shared linear transformation that applies to every node.

The normalized attention coefficients are then used to compute a linear combination of the features to serve as the final output for the nodes after applying a nonlinearity σ:

$$\vec{h}_i' = \sigma\left(\sum_{j\in N_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right)$$

Graph Attention Network version 2 (GATv2) - To extend the popular GAT architecture, GATv2 is proposed to upgrade the static attention to be dynamic depending on the query node [33]. The difference to GAT is that this version applies a layer after the nonlinearity LeakyReLU, which could be observed from below formula:

$$\alpha_{ij} = \frac{\exp\left(\vec{a}^T\cdot\text{LeakyReLU}\left(\left[\mathbf{W}\vec{h}_i \,\|\, \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in N_i}\exp\left(\vec{a}^T\cdot\text{LeakyReLU}\left(\left[\mathbf{W}\vec{h}_i \,\|\, \mathbf{W}\vec{h}_k\right]\right)\right)}$$

Graph Sample and Aggregate (GraphSAGE) uniformly samples a fixed number of neighbors instead of the entire neighborhood. Unlike GCN, this approach uses various aggregation architectures such as Mean aggregator, LSTM aggregator, and Pooling aggregator [34]. In this study, the LSTM aggregator proposed by Hochreiter and Schmidhuber [35] is used. Since LSTM is inherently asymmetric due to its sequential processing of inputs, GraphSAGE is adapted to handle an unordered set of neighbors by applying LSTM to a random permutation of the neighbors.

4. Results

4.1. Baseline method results

The baseline method is only suitable for addressing RQs 1 and 2, as RQ 3 involves a multi-value prediction problem. The application of binary prediction methods to multi-label classification scenarios necessitates a holistic consideration of all labels. Conventionally, the loss function is computed as an aggregation across all labels. However, this approach may not be optimal in the context of medical informatics, particularly in pharmacovigilance. The summation of losses can potentially lead to the oversight of critical ADEs, as the model might prioritize overall loss minimization at the expense of detecting less frequent but clinically significant events. The baseline method analysis focuses on extracting network centrality measures, such as weighted degree centrality, eigenvector centrality, closeness centrality, betweenness centrality, and clustering coefficient, to predict the likelihood of ADEs based on patient-specific features, such as age and gender, as well as network features. With this baseline approach, the GridSearchCV technique from the Scikit-Learn library [36] was used to search for the optimal hyperparameters. The results of the two questions are presented in Table 5.

Answering RQ 1, the Random Forest algorithm demonstrated superior performance, achieving an accuracy of 0.8011 along with better performance across other evaluation metrics. Similarly, in RQ 2, Random Forest was generally the best-performing classifier, exhibiting

Table 5
Results of the machine learning algorithms applied to patients' attributes and network centrality measures for RQ 1 and RQ 2.

RQ 1- classify two cohorts				
Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7473	0.7505	0.7409	0.7456
Naïve Bayes	0.7400	0.7256	0.7719	0.7480
K Nearest Neighbors	0.6943	0.6833	0.7245	0.7033
Support Vector Machine	0.7646	0.7636	0.7664	0.7650
Decision Tree	0.7573	0.7681	0.7372	0.7523
Random Forest	**0.8011**	**0.7936**	**0.8139**	**0.8036**
RQ 2 - predict the timing of ADEs				
Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7414	0.7822	0.6529	0.7117
Naïve Bayes	0.7475	**0.8162**	0.6240	0.7073
K Nearest Neighbors	0.8202	0.7865	0.8678	0.8251
Support Vector Machine	0.7737	0.7802	0.7479	0.7637
Decision Tree	0.8000	0.7719	0.8388	0.8040
Random Forest	**0.8323**	0.7955	**0.8843**	**0.8376**

an accuracy of 0.8323. Of note, Naïve Bayes achieved the highest precision. Nevertheless, the absolute values of these performance metrics are not entirely satisfactory.

4.2. Experimental settings for the proposed model

The datasets for each of the three RQs were randomly divided into training, validation, and test sets based on a ratio of 0.6:0.2:0.2. This ensures enough data for training, validation, and testing while minimizing bias. To train the graph-based models, we utilized the PyTorch Geometric (PyG) library [37]. For all four architectures (GCN, GAT, GATv2Conv, and GraphSAGE), a three-layer neural network was implemented with an input of 128 embedding dimensions and hidden layers with 256, 128, and 64 neurons. The batch size was 256. All the models were trained for a maximum of 500 epochs using the Adam optimizer [38], with early stopping at 20 epochs. The convergence criteria are to let the training run for up to 500 epochs unless it stops early by detecting 20 consecutive epochs where the losses do not change. ReLu was used as the activation function for the hidden layers. The Sigmoid function was used as the activation function for the last layer. The selected parameter ranges are as follows: the dropout rate (p) varies from 0.1 to 0.9 in increments of 0.1, the top-k pooling ratio ranges from 0.2 to 0.8 in increments of 0.05, and the learning rate includes values of 0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5. Binary Cross-Entropy Loss for Q1,2 and BCEWithLogitsLoss for Q3 was employed to measure and minimize the loss during training. After testing various combinations of these relevant hyperparameters, the best ones used in the three RQs are in Table 6.

Table 6
Hyper-parameters of the four models (GCN, GAT, GATv2Conv, GraphSAGE) for the three RQs, respectively.

		RQ 1	RQ 2	RQ 3
GCN	Dropout (p)	0.5	0.5	0.3
	Top k Pooling Ratio	0.8	0.8	0.9
	Learning Rate	0.001	0.001	0.003
GAT	Dropout (p)	0.5	0.5	0.3
	Top k Pooling Ratio	0.8	0.8	0.9
	Learning Rate	0.001	0.001	0.003
GATv2Conv	Dropout (p)	0.5	0.5	0.3
	Top k Pooling Ratio	0.8	0.8	0.9
	Learning Rate	0.001	0.001	0.003
GraphSAGE	Dropout (p)	0.5	0.5	0.3
	Top k Pooling Ratio	0.8	0.8	0.9
	Learning Rate	0.001	0.001	0.003

4.3. Evaluation and discussion

4.3.1. Evaluation metrics

To evaluate the prediction results, we employed five metrics: Accuracy, Precision, Recall, F1-score, and AUROC. These metrics are defined based on the comparison of predicted labels against ground truth labels. The confusion matrix used for these calculations is presented below in Table 7, followed by the definitions of the five metrics.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2^{*}\ Precision^{*}Recall}{Precision + Recall}$$

The calculation of AUROC is based on True Positive Rate (TPR) and False Positive Rate (FPR):

$$TPR = Precision = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

The area under the curve of TPR against FPR according to different thresholds is calculated as Area Under the Receiver Operating Characteristic curve (AUROC) and this is displayed in Fig. 6.

4.3.2. Results from the entire cohort

The results obtained from the four GNN-based models for each RQ are summarized in Table 8. Overall, the four GNN models have yielded satisfactory results based on their accuracy, F1-score, and AUROC scores for all three RQs, and they have outperformed the baseline model in all metrics for RQs 1 and 2. To evaluate the performance differences among the four proposed GNN variants, we employed independent Student's t-tests. This statistical analysis was conducted to determine whether the superior performance of the highest-scoring variant was consistent and statistically significant, rather than occurring by chance. For example the first test would be to compare GraphSAGE and GCN for their accuracy. The null hypothesis is GraphSAGE is not better than GCN in accuracy and the alternative hypothesis is that GraphSAGE does provide better accuracy. Similar tests are all done for each pair of results and the significance threshold is set at 0.05 as is conventionally used in statistics. The results of these comparative analyses are presented in Table 9. This approach allows us to assess the reliability and consistency of the performance advantages observed in the highest-performing GNN variant across multiple trials.

Regarding RQ 1, GraphSAGE exhibited superior performance in all metrics, with an accuracy of 0.8863. Moreover, this model demonstrated remarkable stability, as evidenced by its lowest standard deviation for accuracy and AUROC, which were 0.0146 and 0.0152. This finding can be attributed to the inductive nature of GraphSAGE, which is deemed to be transductive enough to allow for efficient generalization to unseen nodes in evolving graphs, in contrast to other GNN-based algorithms [34]. It is worth mentioning that all models outperformed the baseline

Table 7
Confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

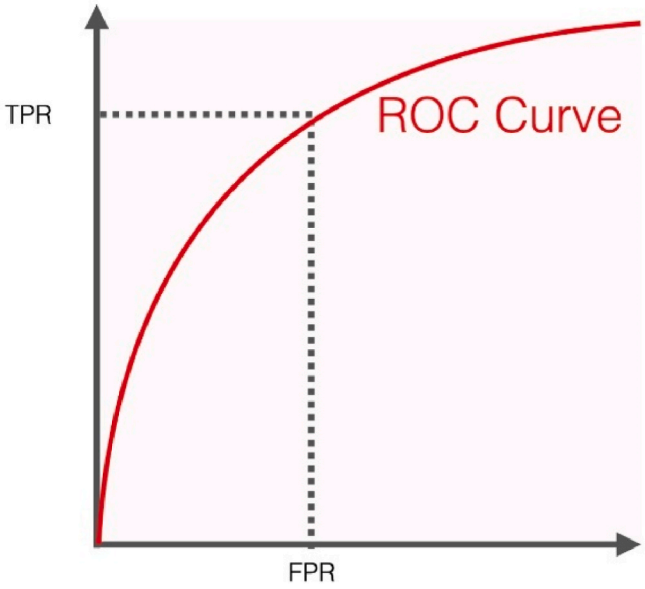

Fig. 6. The definition of AUROC.

Table 8
Results of the four models (GCN, GAT, GATv2Conv, GraphSAGE) for the three RQs.

RQ 1- classify two cohorts					
Models	Accuracy	Precision	Recall	F1-score	AUROC
GCN	0.8631 ± 0.0236	0.8551 ± 0.0275	0.9003 ± 0.0270	0.8769 ± 0.0221	0.8597 ± 0.0232
GAT	0.8717 ± 0.0149	0.8706 ± 0.0241	0.8997 ± 0.0188	0.8841 ± 0.0126	0.8692 ± 0.0161
GATv2Conv	0.8676 ± 0.0206	0.8703 ± 0.0237	0.8944 ± 0.0268	0.8819 ± 0.0188	0.8643 ± 0.0209
GraphSAGE	**0.8863 ± 0.0146**	**0.8812 ± 0.0213**	**0.9128 ± 0.0192**	**0.8965 ± 0.0132**	**0.8841 ± 0.0152**
RQ 2 - predict the timing of ADEs					
Models	Accuracy	Precision	Recall	F1-score	AUROC
GCN	0.8763 ± 0.0177	0.9069 ± 0.0224	0.8355 ± 0.0354	0.8691 ± 0.0194	0.8751 ± 0.0169
GAT	**0.8769 ± 0.0206**	0.9072 ± 0.0249	**0.8402 ± 0.0403**	**0.8718 ± 0.0243**	**0.8760 ± 0.0205**
GATv2Conv	0.8569 ± 0.0280	**0.9103 ± 0.0333**	0.7873 ± 0.0611	0.8423 ± 0.0319	0.8562 ± 0.0271
GraphSAGE	0.8753 ± 0.0257	0.9054 ± 0.0324	0.8328 ± 0.0394	0.8669 ± 0.0267	0.8744 ± 0.0253
RQ 3 - predict what ADE(s) would occur					
Models	Accuracy	Precision	Recall	F1-score	AUROC
GCN	0.9253 ± 0.0051	0.3634 ± 0.0211	0.9781 ± 0.0091	0.5296 ± 0.0277	0.9558 ± 0.0056
GAT	0.9346 ± 0.0054	0.3567 ± 0.0233	0.9764 ± 0.0081	0.5221 ± 0.0255	0.9547 ± 0.0056
GATv2Conv	0.9336 ± 0.0067	0.3547 ± 0.0258	0.9797 ± 0.0066	0.5203 ± 0.0285	0.9558 ± 0.0055
GraphSAGE	**0.9367 ± 0.0045**	**0.3709 ± 0.0156**	**0.9812 ± 0.0076**	**0.5382 ± 0.0165**	**0.9581 ± 0.0048**

model across all metrics, thus underscoring the effectiveness of GNN-based models for ADE prediction.

The motivation for setting up RQ 2 was to investigate whether the ICD codes of non-ADE-associated patients played a crucial role in the predictive power of the models in RQ 1. By not including any negative patients in RQ 2, the models can more explicitly learn from the positive patients' ICD code histories, enabling the prediction of the timing when ADEs might occur. Interestingly, the GAT model was found to be the best-performing model among the four for RQ 2, while GraphSAGE was the best for RQ 1.

In RQ 3, the GraphSAGE model exhibits the highest accuracy of 0.9367, while all other models yield results above 92 %. Notably, the standard deviation for RQ 3 is remarkably small across various experiments. With respect to accuracy measurement, GraphSAGE's standard deviation is merely 0.0045, denoting that the model is stable and robust. In comparison to RQs 1 and 2, the low precision of RQ 3 is comprehensible and credible. The GraphSAGE model's precision fluctuates around 0.3709, signifying that only about 37 % of predicted positives correspond to the actual values. Nevertheless, considering the Recall's high value of 98.12 %, it can be inferred that GraphSAGE correctly predicts 98.12 % of actual positive subgraphs despite low precision. It is noteworthy that in this medical setting, a higher recall is more valuable, as detecting possible ADEs provides greater assurance of patient safety, even if it results in some false positive cases. The low precision also implies that they could be used to recognize possible unreported ADEs. The global ADE research area has encountered a contentious issue of underreporting, as is also the case in Australia [39]. Therefore, these models could predict unobserved ADEs, thereby serving as a tool to address the possible underreporting issue and providing additional confidence to researchers and patients. The primary clinical utility would be the capability to forecast the occurrence and identify the type of ADEs during a patient's admission, leveraging their past admission history. This predictive insight would enable clinicians to proactively intervene, thereby mitigating the associated risks. Taking a step back to address the high recall but low precision issue, implementing rule-based filters or secondary machine learning models to refine the initial predictions would be a suitable approach in this medical context.

The results from the independent t-tests presented in Table 9 examine the null hypothesis that the best model for each RQ does not have a significantly greater performance metric score compared to the other models, and the alternative hypothesis is that the best model has a higher performance score than the corresponding compared models, which assesses that the result is not due to chance. For RQ 1, the null hypothesis is rejected for all performance metrics except Recall when comparing GraphSAGE to GCN and GAT. In these cases, GraphSAGE exhibits a significantly better performance than the other models. We also note that the p-value in the failed tests is just slightly greater than 0.05. In RQ 2, the null hypothesis is rejected only when comparing the best model with GATv2Conv. That means only GATv2Conv's performance is significantly lower than all the other three. For RQ 3, the null hypothesis is rejected for precision and recall when comparing GraphSAGE to GAT and GATv2Conv. Overall, these results suggest that GraphSAGE and GAT generally outperform other models in their respective best-performing RQs. Further analysis or a larger sample size may be needed to determine whether these results hold up or if additional factors might influence model performance.

We have also added a table to show the time complexity of the proposed model in Table 10 where the execution time of various graph neural network models was compared across three RQs. The GAT model exhibited the highest variability, particularly in RQ 1, with a standard deviation of 886.85 s. GATv2Conv demonstrated the longest execution time for RQ 2 at an average of 4584.47 s. Interestingly, GraphSage performed the fastest for RQ 3. Overall, the choice of model significantly impacts execution time, with performance varying across different RQs.

4.3.3. Results from ADE segments

Further, experiments were conducted focusing on ADE segments, delineated according to the classification system proposed by Hohl and the detailed description of these categories has been presented in the data preparation section. These further experimentations were conducted utilizing the GraphSAGE framework, focusing on a subset of the ADEs delineated in the accompanying table. These ADEs were categorized into five subsections: A, B, C, D, and E. This segmentation was pursued with the objective of exploring whether distinct identification

Table 9
Independent *t*-test examining the null hypothesis that the best model for each question does not have a significantly greater performance metric score than others.

RQ 1- classify two cohorts					
	Accuracy	Precision	Recall	F1-score	AUROC
	GraphSAGE	GraphSAGE	GraphSAGE	GraphSAGE	GraphSAGE
GCN	0.0004	0.0009	0.0637	0.0003	0.0002
GAT	0.0002	0.0908	0.0002	0.0004	0.0000
GATv2Conv	0.0000	0.0192	0.0003	0.0000	0.0000
RQ 2 - predict the timing of ADEs					
	Accuracy	Precision	Recall	F1-score	AUROC
	GAT	GATv2Conv	GAT	GAT	GAT
GCN	0.2218	0.1659	0.3295	0.3110	0.2385
GAT	–	0.2972	–	–	–
GATv2Conv	0.0025	–	0.0005	0.0001	0.0003
GraphSAGE	0.1489	0.2550	0.1961	0.1272	0.1984
RQ 3 - predict what ADE(s) would occur					
	Accuracy	Precision	Recall	F1-score	AUROC
	GraphSAGE	GraphSAGE	GraphSAGE	GraphSAGE	GraphSAGE
GCN	0.4235	0.7747	0.4669	0.5458	0.8837
GAT	0.8619	0.0268	0.0495	0.7797	0.5625
GATv2Conv	0.5189	0.0049	0.0495	0.1788	0.8035

Table 10
Execution time comparison of graph neural network models across RQs (in seconds).

	RQ1	RQ2	RQ3
GCN	1301.13 ± 34.77	2113.62 ± 131.30	1676.40 ± 57.86
GAT	2712.66 ± 886.85	3124.85 ± 151.51	1422.30 ± 30.93
GATv2Conv	1879.52 ± 56.98	4584.47 ± 353.87	1641.70 ± 56.65
GraphSage	1865.91 ± 110.25	2518.12 ± 108.40	1391.76 ± 46.88

of causal relationships within these subsections could yield varied outcomes. The results of these experiments are presented in Table 11.

Upon examining the results by subsection, it appears that RQs 1 and 3 do not exhibit significant variance across the different categories; they have all maintained a nearly identical level of predictive power and robustness. However, it is notable that the outcomes from RQ 2 reveal that subsections B, C, and E demonstrate a much higher level of accuracy (0.9144, 0.9031, and 0.9271, respectively) compared to the general results (0.8753), while subsections A and D exhibit considerably lower accuracy (0.7231 and 0.7144, respectively). This warrants further examination to explore potential causes, such as the spatial distribution of various sections or the unique intrinsic characteristics that distinguish the different categories as they have been identified based on different key words or different causal ratings.

Interpreting how graph-based models predict and classify different cohorts has been a challenging task due to the fact that latent information present in the hidden layers is difficult to visualize [40]. We utilized the final convolution layer of the four models for RQ 1 and plotted their abilities to distinguish between the two cohorts in Fig. 7a using the Seaborn library [41]. The yellow circles represent positive patients, while the blue circles represent negative patients. As the dimension of the last hidden layer is 64, principal component analysis (PCA) was applied to the convolution vector in the last hidden layer, and a three-dimensional vector was used to make the plots. It is apparent that GraphSAGE and GAT are better at classifying the two classes of data points, but it is not visually discernible which one between these two is superior. The same plotting technique was also used for RQ 2 in Fig. 7b.

The practical implementation of this model represents a significant advancement in pharmacovigilance and patient safety. The current system for detecting ADEs primarily relies on passive reporting mechanisms, wherein clinicians and patients voluntarily report incidents after

Table 11
Results of sub-sections using GraphSAGE setups.

RQ 1- classify two cohorts					
Sub-sections	Accuracy	Precision	Recall	F1-score	AUROC
A	0.8859 ± 0.0183	0.8851 ± 0.0213	0.9045 ± 0.0209	0.8946 ± 0.0180	0.8841 ± 0.0180
B	0.8829 ± 0.0163	0.8808 ± 0.0242	0.9040 ± 0.0180	0.8920 ± 0.0164	0.8814 ± 0.0170
C	0.8894 ± 0.0158	0.8797 ± 0.0220	0.9232 ± 0.0197	0.9006 ± 0.0142	0.8858 ± 0.0166
D	0.8866 ± 0.0141	0.8757 ± 0.0243	0.9209 ± 0.0193	0.8974 ± 0.0127	0.8838 ± 0.0148
E	0.8846 ± 0.0188	0.8851 ± 0.0206	0.9049 ± 0.0216	0.8947 ± 0.0166	0.8829 ± 0.0191
RQ 2 - predict the timing of ADEs					
Sub-sections	Accuracy	Precision	Recall	F1-score	AUROC
A	0.7231 ± 0.0386	0.7563 ± 0.0592	0.6811 ± 0.0585	0.7138 ± 0.0367	0.7254 ± 0.0385
B	0.9144 ± 0.0215	0.9122 ± 0.0341	0.9194 ± 0.0315	0.9151 ± 0.0208	0.9144 ± 0.0205
C	0.9031 ± 0.0326	0.9120 ± 0.0297	0.8919 ± 0.0416	0.9015 ± 0.0322	0.9034 ± 0.0314
D	0.7144 ± 0.0390	0.7291 ± 0.0657	0.6896 ± 0.0434	0.7067 ± 0.0374	0.7158 ± 0.0394
E	0.9271 ± 0.0123	0.9454 ± 0.0337	0.9041 ± 0.0180	0.9237 ± 0.0144	0.9267 ± 0.0125
RQ 3 - predict what ADE(s) would occur					
Sub-sections	Accuracy	Precision	Recall	F1-score	AUROC
A	0.9378 ± 0.0029	0.3711 ± 0.0122	0.9809 ± 0.0099	0.5384 ± 0.0135	0.9585 ± 0.0058
B	0.9359 ± 0.0049	0.3625 ± 0.0258	0.9794 ± 0.0045	0.5286 ± 0.0275	0.9568 ± 0.0034
C	0.9370 ± 0.0046	0.3634 ± 0.0200	0.9811 ± 0.0050	0.5301 ± 0.0217	0.9582 ± 0.0034
D	0.9386 ± 0.0044	0.3710 ± 0.0194	0.9817 ± 0.0054	0.5382 ± 0.0207	0.9593 ± 0.0041
E	0.9387 ± 0.0044	0.3680 ± 0.0180	0.9847 ± 0.0078	0.5355 ± 0.0186	0.9608 ± 0.0022

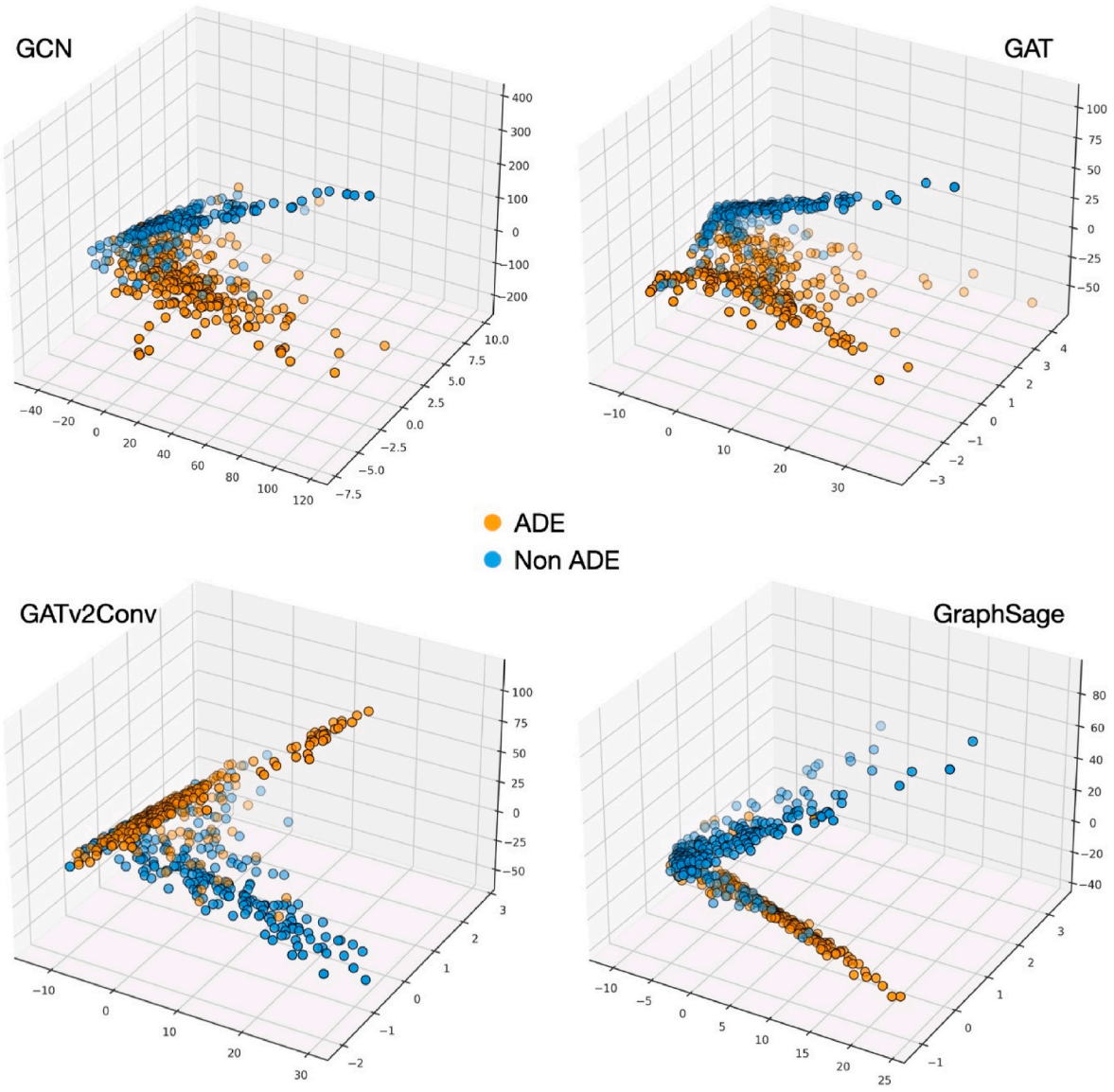

Fig. 7a. Three-dimensional plots of the final convolution layer in the four proposed models for RQ 1.

they occur [42]. This approach, while valuable, is inherently limited by substantial underreporting, which can lead to delayed recognition of potential drug safety issues and inadequate risk assessment. By leveraging administrative data, as demonstrated in our study, we have developed a more proactive and comprehensive approach to ADE detection and prediction, which is critical to improving the detection rate and ability to predict ARS at a given presentation.

Regardless, this study has several limitations. First, any window time between diseases were not considered, which means a month's gap between diseases or a year's gap between diseases were embedded with the same methods. As a result, ICD codes associated with an ADE could only serve as a connection in the entire graph without providing further insights into their relationships with a specific ADE. Second, due to people making bulk claims for a couple of prior medication expenses on a single date, the claims data may not accurately reflect ICD-code sequencing, and we did not have the ability to distinguish between primary ICD-code and supportive ICD-code for each admission. Therefore, the sequenced modeling of the ICD codes in the dataset may not reflect reality precisely. Third, personal attributes such as age and gender, which have been shown to be important in ADE studies [43,44], were not considered. Fourth, The results from our method are not compared to other existing works in the literature due to different problem formulations and our attempts to predict various categories of ADEs.

As can be seen from Table 10, training GNNs on large-scale health claims data presents significant computational challenges due to the sheer volume and complexity of the data, let alone that we have only included a small number of patients used as training and testing samples. To manage the challenges effectively, we can employ various strategies such as graph sampling techniques, distributed computing, and GPU acceleration. For instance, node-wise or subgraph sampling can reduce memory requirements, while distributed processing across multiple machines can handle larger datasets. To ensure scalability, incremental learning methods and model compression techniques like knowledge distillation can be utilized. By combining these approaches, it can be expected to develop GNN models that are both computationally efficient and capable of extracting valuable insights from vast health claims datasets.

5. Conclusion

This study proposes a novel approach to predict ADEs using a

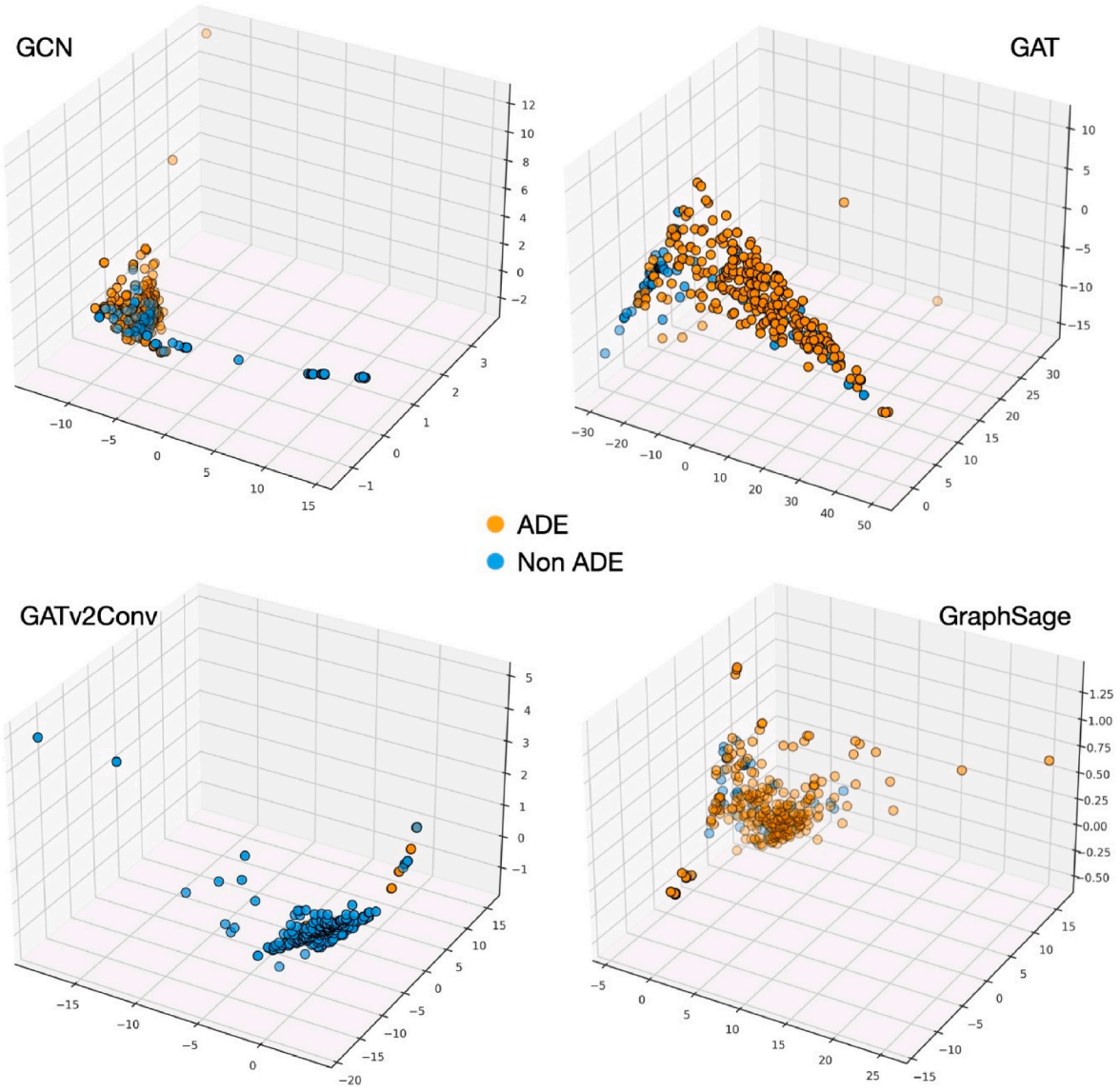

Fig. 7b. Three-dimensional plots of the final convolution layer in the four proposed models for RQ 2.

network of ICD-10 codes extracted from CBHS claims history. By modeling patients as subgraphs and applying GNN-based machine learning methods to these subgraphs, the study aims to answer three RQs: the likelihood of a patient experiencing an ADE, the timing of such events, and the specific ADEs a patient might develop. Experimental results showed that GraphSAGE had the highest accuracy for RQs 1 and 3, 0.8863 and 0.9367, respectively, while GAT had the best performance for RQ 2, which was 0.8769. These findings highlight the potential of GNNs to accurately model the complex relationships between ICD-10 codes and predict the occurrence of ADEs. These models could be useful in drug development and clinical settings to improve patient care. In light of the limitations identified in this study, several avenues for future research can be considered to enhance the framework and further contribute to the field of adverse drug event prediction. Future studies could explore the integration of time-based information [45], capturing the time gaps between diseases and leveraging them to gain deeper insights into the relationships between ICD codes and specific ADEs. Additionally, finding methods to distinguish primary and supportive ICD codes within the claims data would further refine the accuracy and real-world applicability of the model. To address scalability challenges associated with larger datasets, we could investigate the integration of graph machine learning with federated learning approaches [46] or consider a fog computing environment [47], which could potentially reduce computational time and enhance model performance.

CRediT authorship contribution statement

Fangyu Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matloob Khushi:** Writing – review & editing, Supervision. **Jonathan Brett:** Writing – review & editing. **Shahadat Uddin:** Writing – review & editing, Validation, Supervision.

Data availability

This study obtained research data from an Australian private health insurance organization (Commonwealth Bank Health Society, CBHS). This data was collected in a de-identified format and through a research agreement between the CBHS and the University of Sydney (University of Sydney reference number: CT18435). For reproducing the results of this study, the relevant data and codes of the study can be accessed from this repository: https://doi.org/10.5281/zenodo.7703238.

Declaration of competing interest

None Declared.

Acknowledgment

MK is supported by UKRI NERC grant NE /X000192/12.

References

[1] K. Yu, R.L. Nation, M.J. Dooley, Multiplicity of medication safety terms, definitions and functional meanings: when is enough enough? BMJ Qual. Saf. 14 (5) (2005) 358–363.
[2] N. El Morabet, et al., Prevalence and preventability of drug-related hospital readmissions: a systematic review, J. Am. Geriatr. Soc. 66 (3) (2018) 602–608.
[3] R. Lim, et al., The extent of medication-related hospital admissions in Australia: a review from 1988 to 2021, Drug Saf. 45 (3) (2022) 249–257.
[4] M.A. Hadi, et al., Pharmacovigilance: pharmacists' perspective on spontaneous adverse drug reaction reporting, Integrated Pharm. Res. Pract. 6 (2017) 91–98.
[5] D. Formica, et al., The economic burden of preventable adverse drug reactions: a systematic review of observational studies, Expet Opin. Drug Saf. 17 (7) (2018) 681–695.
[6] The Importance of Pharmacovigilance, World Health Organization, 2002.
[7] T. Morimoto, et al., Adverse drug events and medication errors: detection and classification methods, BMJ Qual. Saf. 13 (4) (2004) 306–314.
[8] C.M. Hohl, et al., ICD-10 codes used to identify adverse drug events in administrative data: a systematic review, J. Am. Med. Inf. Assoc. 21 (3) (2014) 547–557.
[9] A.D. Gholap, et al., Advances in artificial intelligence in drug delivery and development: a comprehensive review, Comput. Biol. Med. (2024) 108702.
[10] E. Pauwels, V. Stoven, Y. Yamanishi, Predicting drug side-effect profiles: a chemical fragment-based approach, BMC Bioinf. 12 (1) (2011) 1–13.
[11] M. Liu, et al., Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs, J. Am. Med. Inf. Assoc. 19 (e1) (2012) e28–e35.
[12] L.-C. Huang, X. Wu, J.Y. Chen, Predicting adverse side effects of drugs, BMC Genom. 12 (5) (2011) 1–10.
[13] R.R. Saxena, R. Saxena, Applying graph neural networks in pharmacology, Authorea Preprints (2024).
[14] D.M. Bean, et al., Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records, Sci. Rep. 7 (1) (2017).
[15] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, Bioinformatics 34 (13) (2018) i457–i466.
[16] A. Deac, et al., Drug-drug adverse effect prediction with graph co-attention, arXiv preprint arXiv:1905.00534 (2019).
[17] L. Yu, et al., MSDSE: predicting drug-side effects based on multi-scale features and deep multi-structure neural network, Comput. Biol. Med. 169 (2024) 107812.
[18] S. Chen, et al., An effective framework for predicting drug–drug interactions based on molecular substructures and knowledge graph neural network, Comput. Biol. Med. 169 (2024) 107900.
[19] A. Khan, U. Srinivasan, S. Uddin, Development and exploration of polymedication network from pharmaceutical and medicare benefits scheme data, in: Proceedings of the Australasian Computer Science Week Multiconference, Association for Computing Machinery, Sydney, NSW, Australia, 2019. Article 34.
[20] Y. Wang, Z. Yang, Q. Yao, Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning, Commun. Med. 4 (1) (2024) 59.
[21] D.A. Nguyen, C.H. Nguyen, H. Mamitsuka, A survey on adverse drug reaction studies: data, tasks and machine learning methods, Briefings Bioinf. 22 (1) (2021) 164–177.
[22] F. Zhou, S. Uddin, Mining adverse drug events from patients' disease histories via a GNN-based subgraph prediction method, in: Proceedings of the Australasian Computer Science Week Multiconference, 2023.
[23] H. Luo, et al., Drug-drug interactions prediction based on deep learning and knowledge graph: a review, iScience 27 (3) (2024) 109148.
[24] CBHS Health, Commonwealth Bank health society. www.cbhs.com.au, 2022.
[25] L. Zhou, A.P. Rupa, Categorization and association analysis of risk factors for adverse drug events, Eur. J. Clin. Pharmacol. 74 (4) (2018) 389–404.
[26] T.B. Agbabiaka, J. Savović, E. Ernst, Methods for causality assessment of adverse drug reactions, Drug Saf. 31 (1) (2008) 21–37.
[27] A. Khan, S. Uddin, U. Srinivasan, Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression, Int. J. Med. Inf. 115 (2018) 1–9.
[28] A. Khan, S. Uddin, U. Srinivasan, Chronic disease prediction using administrative data and graph theory: the case of type 2 diabetes, Expert Syst. Appl. 136 (2019) 230–241.
[29] H. Lu, et al., A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus, Appl. Intell. 52 (3) (2022) 2411–2422.
[30] F. Zhou, S. Uddin, Interpretable drug-to-drug network features for predicting adverse drug reactions, Healthcare 11 (4) (2023) 610.
[31] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).
[32] P. Velickovic, et al., Graph attention networks, Stat 1050 (2017) 20.
[33] S. Brody, U. Alon, E. Yahav, How Attentive Are Graph Attention Networks?, 2021 arXiv preprint arXiv:2105.14491.
[34] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).
[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[36] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[37] J.E.L. Matthias Fey, Fast Graph Representation Learning with PyTorch Geometric, 2019.
[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[39] G.C. Miller, H.C. Britt, L. Valenti, Adverse drug events in general practice patients in Australia, Med. J. Aust. 184 (7) (2006) 321–324.
[40] H. Yuan, et al., Explainability in graph neural networks: a taxonomic survey, arXiv preprint arXiv:2012.15445 (2020).
[41] M.L. Waskom, Seaborn: statistical data visualization, J. Open Source Softw. 6 (60) (2021) 3021.
[42] G.J. Dal Pan, Ongoing challenges in pharmacovigilance, Drug Saf. 37 (2014) 1–8.
[43] M.B. Zazzara, et al., Adverse drug reactions in older adults: a narrative review of the literature, European geriatric medicine 12 (3) (2021) 463–473.
[44] Y. Yu, et al., Systematic analysis of adverse event reports for sex differences in adverse drug events, Sci. Rep. 6 (1) (2016) 1–9.
[45] H. Chen, H. Eldardiry, Graph time-series modeling in deep learning: a survey, ACM Trans. Knowl. Discov. Data 18 (5) (2024). Article 119.
[46] A. Shahidinejad, et al., Context-aware multi-user offloading in mobile edge computing: a federated learning-based approach, J. Grid Comput. 19 (2) (2021) 18.
[47] M. Salimian, M. Ghobaei-Arani, A. Shahidinejad, Toward an autonomic approach for Internet of Things service placement using gray wolf optimization in the fog computing environment, Software Pract. Ex. 51 (8) (2021) 1745–1772.