GraphEHR: Heterogeneous Graph Neural Network for Electronic Health Records

Kushagra Agarwal* Minyoung Choe* Nivedhitha Dhanasekaran* Juho Jung*
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{kagarwa2, mchoe2, ndhanase, juhoj}@andrew.cmu.edu

Abstract

This project introduces a new approach to healthcare analytics by leveraging Heterogeneous Graph Neural Networks (HGNNs) for multiple predictive tasks using Electronic Health Records (EHR) data. GraphEHR effectively uncovers intricate relationships among medical concepts, demonstrating robustness in understanding diverse graph structures. We used the MIMIC-III EHR database to benchmark performance improvements across 13 predictive tasks, including mortality, advanced cancer, and depression. The significance of this work lies in its contribution to computational biomedicine research and healthcare analytics, where the vast information within EHRs is harnessed to drive patient care. We compare the performance of GraphEHR against multiple baseline models across the different tasks, showcasing its ability to comprehend the complexity and multi-dimensional nature of EHR data. This work represents a transformative step towards harnessing the full potential of EHR data through advanced graph neural network approaches, ultimately impacting healthcare decision-making and improving patient outcomes.

1 Introduction

In the landscape of healthcare analytics, the pivotal role played by Electronic Health Records (EHR) cannot be overstated. These vast patient data repositories harbor the potential to revolutionize healthcare outcomes, provided that their intricate nature is effectively navigated. This study addresses the imperative task of unraveling latent patterns within EHR data, specifically focusing on tasks crucial for patient prognoses, such as mortality prediction [19].

Recognizing the complexity and multidimensional structure inherent in EHR datasets, we propose a Heterogeneous Graph Neural Network (HGNN), GraphEHR, tailored to address diverse predictive tasks. It leverages the power of graph neural networks to learn intricate interconnections among medical concepts within the intricate tapestry of EHR data. By introducing a heterogeneous graph network, our approach exhibits robustness in learning graph structures, translating to adaptive enhancements in performance across a spectrum of EHR predictive tasks.

In the United States alone, a staggering number of nearly 795,000 individuals face permanent disability or death annually due to misdiagnosis, incurring a substantial economic cost of approximately $20 billion each year [26]. This underscores the critical need for precise patient diagnosis, where machine learning collaborates with medical professionals to bridge the gap. Our project, GraphEHR, takes a crucial stride in this direction. The data generated during a patient's hospital stay is diverse, comprising medication information for both solid and liquid drugs, monitoring data for vital blood markers, and procedural data related to surgeries. We harness this multi-modal information to predict diagnoses accurately. However, a key challenge arises when a patient experiencing symptoms such as

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

severe abdominal pain visits the hospital multiple times, each time receiving treatment for a specific symptom without a holistic understanding of their medical history.

GraphEHR addresses this challenge by integrating data from multiple hospital visits and drawing insights from Electronic Health Records. For instance, as illustrated in Figure 1, a patient initially treated for abdominal pain and later admitted with jaundice may prompt suspicion of an underlying condition. By analyzing the integrated records, our system has the potential to predict the early stages of diseases like pancreatic cancer, which often present with similar initial symptoms.

Significance: The implications of such a diagnostic system are profound, bringing about life-changing benefits for patients and their families. It enables more accurate and timely diagnoses and translates into substantial cost savings for insurance companies by mitigating the expenses associated with misdiagnosis. Furthermore, doctors gain the ability to meaningfully impact more lives, contributing to a healthcare landscape where informed decisions lead to improved patient outcomes. Our work holds profound significance in the realm of healthcare analytics, where the impact on patient outcomes is direct and transformative. Electronic Health Records (EHR) serve as reservoirs of invaluable information, offering a pathway to enhance the precision of diagnosis, treatment, and overall patient care. The integration of Graph Neural Networks (GNNs) in our approach represents a pioneering step toward unlocking new possibilities in accurate and insightful healthcare applications.

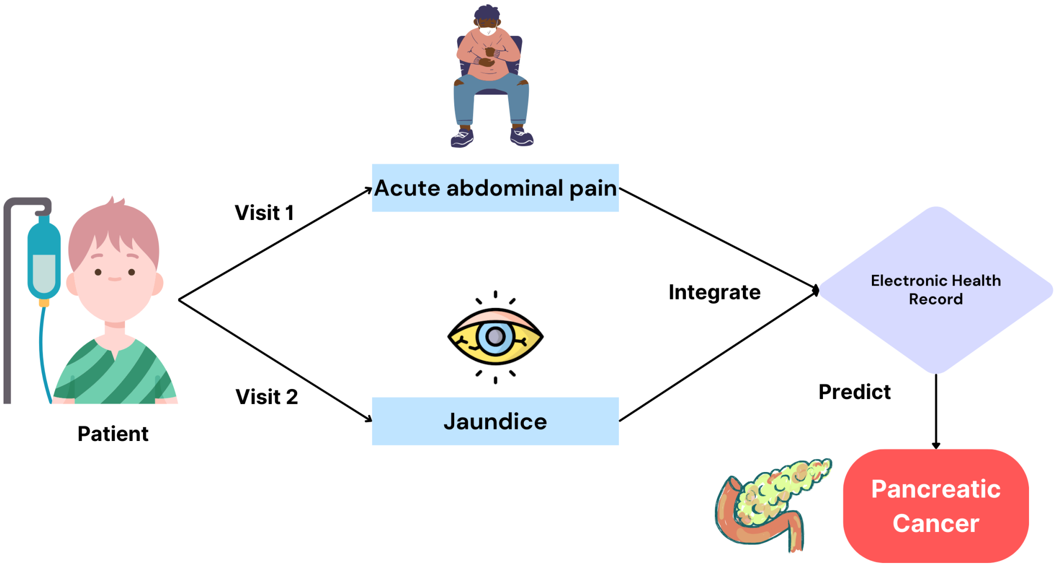

Figure 1: This figure demonstrates how utilizing the rich data in EHR can improve the accuracy of diagnosis. Let's say you experience severe abdominal pain one day and decide to visit the hospital. The doctor will treat you with painkillers and send you back. After some days, you are admitted again with jaundice. The doctor, unsuspecting that it could be a sign of something else, treats you for jaundice and sends you back. We need to integrate data from multiple visits to improve diagnosis, which can be done by looking at the complete Electronic Health Record. Using GraphEHR on top of this integrated record, we might be able to predict the early stages of Pancreatic cancer, which has exactly the same initial symptoms.

Our study extends the work on graphical learning techniques studied by Zhu et al., [29], by exploring Heterogeneous Graph Neural Networks for the MIMIC-III [14] EHR dataset. We hypothesize that incorporating a heterogeneous graph neural network technique significantly augments the model's capacity to handle diverse information types effectively. This augmentation could lead to an effective representation of the complex relationships inherent in EHR data. To verify our hypotheses, the proposed Heterogeneous Graph Neural Network, **GraphEHR**, is evaluated for several downstream prediction tasks like mortality prediction, advanced cancer, advanced heart disease etc.

In the subsequent sections of this report, we provide a comprehensive exploration of our methodology. Section 2 delves into the prior literature, setting the stage for our research. Section 3 offers a concise yet thorough overview of the GraphEHR model. The detailed exposition of our model's architecture and extensions to the base paper's methodology is presented in Section 4. Moving forward, Section

5 elucidates our dataset and the pre-processing paradigm. Sections 6 and 7 describe the evaluation metric and loss function used for model training. Following that, Section 8 encompasses the baseline selection, including the state-of-the-art and other baselines we benchmark against. Section 9 discusses the implementational details for the baselines, while our experiments are cataloged in Section 10. Most importantly, the results are discussed in Section 11. Section 12 explores potential future directions arising from identified limitations in our project. Finally, Section 13 provides a concluding reflection, encapsulating key takeaways, emphasizing significance, delving into future potential, and highlighting our project's broader impact in healthcare analytics and deep learning.

2 Literature Review

In the last decade, Electronic Heath Records (EHR) have become an increasingly popular data source for clinical research and computational biomedicine since they provide the opportunity to enhance patient care through computer program automation. EHRs are maintained by the healthcare provider or its administration to electronically encode the full medical history of a patient, such as demographics, doctor notes, medications, vital signs, immunizations, laboratory test results, radiology reports, etc., as time series data by chronologically recording the patient's visits for their health needs [10]. They were originally intended to be a central repository for storing patient information comprehensively and facilitating the secure transfer of patient medical history between healthcare providers. Over time, with improvements in dealing with privacy and security issues related to EHR data and advancements in data-centric algorithms, they have become the primary source of information for interdisciplinary data-driven clinical studies like disease prediction [11] conducted through collaborations between medical professionals and technologists.

One such well-known, publicly available EHR dataset is MIMIC-III, which stands for Medical Information Mart for Intensive Care 3 [14]. It stores 46520 anonymized (de-identified) medical history records of patients between 2001 and 2012 when they were admitted to the critical care units of Beth Israel Deaconess Medical Center, Boston. It includes 26 tables comprising general patient information, medications, lab tests, clinical notes, vital sign measurements, procedures, etc. A mandatory training course must be completed to access the dataset, and an application requesting access to the dataset must be submitted on the PhysioNet website [14].

Several studies have applied deep learning methodologies to MIMIC-III data [13, 22, 15]. More recently, the two prevalent research tracks in applying deep learning to MIMIC-III are mining for temporal patterns due to the intrinsic time series characteristics of EHR data and learning embeddings of medical concepts without directly modeling time [23, 29].

The former involves collecting features from chronological, biological indicators using deep representation methods like recurrent neural networks [3, 5, 18], convolutional blocks [4, 21], or attention mechanisms [17, 24]. Using representation learning methods, the latter seeks to train deep neural networks that can effectively transform rich medical data into high-dimensional embeddings. One such example is Med2Vec [5], which utilizes skip-gram to learn patient embeddings based on individual patient features. However, the complexity of EHR data goes beyond simple feature embeddings. Notably, diseases can have various causes, and some diseases may co-occur or even lead to the development of other conditions.

The incorporation of graph structures becomes invaluable in capturing this intricate information. Graph-based approaches, such as Graph Convolutional Neural Networks and Graph Attention Neural Networks (e.g., GRAM [5, 6] and MiME), have proven effective in this context. Nevertheless, real-world EHR data often suffer from missing data, and the hierarchical relationships between different features may not be precisely defined.

Graph Convolutional Networks (GCN) [2, 9] are more flexible in learning graph representations than earlier models and can generalize translation-invariant convolution filters in ordinary convolutional neural networks (CNN) to a non-Euclidean localized filter that may be applied to a variety of non-grid input. GCN can be used to learn representations of node features and graph structure through semi-supervised learning for node classification [16]. The model proposed by the paper [29] describes a method for generating labels for unknown nodes based on graph architecture and node attributes. Self-attention, which is comparable to CNN in encoding characteristics from spatial or sequential data [8] takes less computation time and outperforms CNN in language and vision tasks [20, 27]. Graph Attention Network (GAT) [28], like substituting the convolutional block with self-attention, attends each node in the graph on its neighboring nodes and itself to learn localized characteristics instead of utilizing GCN spectrum filters. GAT can assign varied priority to edges in this architecture, boosting model capacity and interoperability and learning graph structures via attention parameters.

The authors of [29] propose an encoder-decoder GNN inspired by the self-attention mechanism and GAT [28] by treating each observed EHR code as a node, first imposing a fully linked graph on them and implicitly learning their graph structure via the self-attention mechanism. Graph Convolution Transformer (GCT) [7], a recent work, presents a visit embedding for each patient medical encounter and connects it with the other medical concept embeddings via the Transformer [27]. Furthermore, the authors address a point raised in their paper: the Transformer's inability to learn attention parameters from scratch, resulting in uniformly distributed attention weights among medical concepts. GCT in [7] overcomes the problem by guiding regularisation with a pre-defined graph.

We build upon the work of [29] and address their research gap of not including Heterogeneous Graph Neural Networks as part of their baseline models. Our rationale was that the provided graph structure already contained sufficient information, and we believe there are more appropriate ways to leverage this existing structure instead of constructing an entirely new one. They employ a graph neural network over the graphs with a substantial variety of information types. For instance, this encompassed relationships like "causes" between A and B and "symptoms" between A and B. This heterogeneity in the graph's information proved confusing for the model during training. To counter this, we adopt a heterogeneous graph neural network approach, which involves creating separate graph neural networks for each edge type and aggregating the resulting embeddings. This strategy aims to enhance the model's ability to effectively handle the diverse information types present in the graph. Finally, using the learned representations, we will perform downstream prediction tasks such as mortality prediction, ICU readmission, phenotype classification, etc.

3 Model Description

Model Description Within our model, we operate on a heterogeneous graph characterized by multiple types of nodes and edges. Specifically, the availability of medication, procedures, and lab results (e.g., heart rate or body temperature) for each patient allows us to establish distinct connections between patients and their respective medications, procedures, or lab results. Our overarching objective is to leverage these inherent relationships between patients. For example, if two patients showcase a history of similar surgical procedures, we can reasonably infer that they will likely suffer from a similar condition.

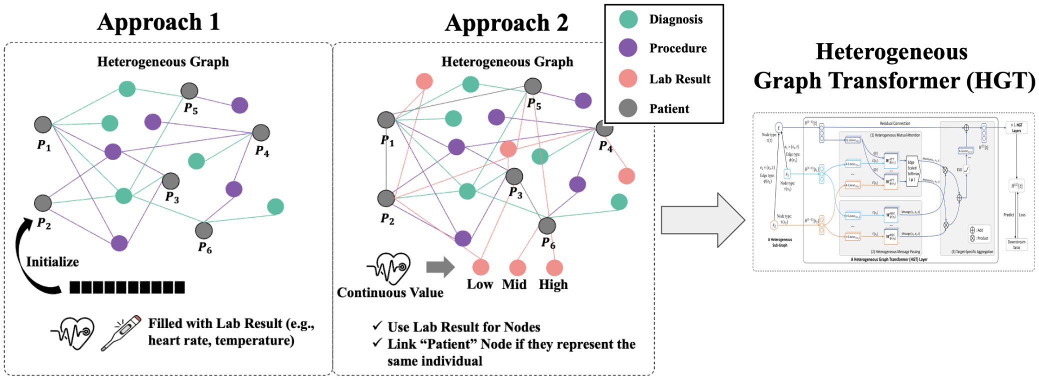

Figure 2: An overall architecture of the proposed method. The proposed heterogeneous graph identifies Electronic Health Record (EHR) features as distinct nodes, each connected by different edges to represent their relationships. The Graph Attention Transformer comprehensively understands the connections between these nodes and all patients, utilizing node embeddings to formulate representations. Finally, a task-specific representation is generated through the fusion of each node representation.

To construct heterogeneous graphs for our study, we employed two distinct approaches. In the *first* approach, we defined three types of nodes: diagnoses, procedures, and patients. Two types of edges were utilized to represent (1) "the patient's diagnosis" and (2) "the patient's procedures." The initialization of patient node features draws from lab results, with the dimension of the patient feature aligning with the number of distinct types and each value corresponding to the patient's specific lab results. We used an embedding layer for the initialization of diagnoses and procedures features.

In the *second* approach, an additional type of node and two additional types of edges were introduced compared to the first approach. Continuous values, such as lab results, were categorized into bins (high, medium, and low). This discretization transforms continuous values into discrete categories, enabling handling lab results as distinct nodes. In this context, we also established connections signifying "the patient's lab results." Furthermore, acknowledging that a patient may have multiple hospital visits, resulting in the existence of more than one node for the same patient, we establish connections between identical patient admission nodes.

In both approaches, a Heterogeneous Graph Transformer (HGT) [12] is used to derive node embeddings from the heterogeneous graphs. The model incorporates distinct embedding layers for each node type and utilizes multi-head attention for neighboring nodes, considering their types and edge types. The ultimate prediction of labels is accomplished through a linear layer applied to the patient node embeddings. This comprehensive framework facilitates the nuanced analysis of patient relationships and contributes to accurate predictions in our heterogeneous medical graph setting.

4 Proposed Extensions

We build upon the foundation laid by the base paper, [29], focusing primarily on addressing the research gap related to the absence of Heterogeneous Graph Neural Networks (HGNN) in their baseline models, improving their embedding structures and extending the model to handling several downstream predictive tasks. Recognizing that the provided graph structure embeddings already carry substantial information, we advocate for leveraging this existing structure more effectively rather than constructing an entirely new one. The base paper, [29], utilized a graph neural network with diverse information types, including relationships such as "causes" between A and B and "symptoms" between A and B. However, this form of heterogeneity posed challenges during model training, leading to confusion at both development and prediction [29]. Furthermore, we would like our HGNN to be versatile and adaptive to several prediction tasks to fully utilize the pool of patient information in EHR data.

To overcome this challenge, our proposed extension involves the following key ideas:

4.1 Enhanced Information Handling with HGNN

We hoped to harness the inherent information richness in graph embedding structures for multi-dimensional EHR data. To this end, we adopt a Heterogeneous Graph Neural Network (HGNN) approach, creating a graph neural network for that can aggregate the information into embeddings that capture and utilize the diverse information types in EHR data effectively. We expected this to improve the model's ability to handle complex relationships by tailoring the learning process to specific edge types, enhancing overall model performance.

4.2 Downstream Predictive Tasks

In order to extend the model's utility beyond the graph structure by applying the learned representations to critical healthcare predictive tasks, we utilized the enhanced HGNN model for downstream prediction tasks such as mortality prediction, advanced cancer prediction etc. The results showcase the versatility and practical applicability of the proposed approach to addressing real-world healthcare challenges.

The proposed extensions aim to address the identified gap in [29] and elevate the model's performance and applicability in healthcare settings through improved information handling, enhanced representational embeddings, and robustness to real-world data challenges.

5 Dataset

We used the MIMIC-III dataset, which contains data for 53,423 distinct hospital admissions for adult patients between 2001 and 2012. If a patient (Subject ID) visits the emergency room (ICU stay ID) multiple times at the same hospital (HADM ID), multiple ICU stay IDs are assigned. Subject ID, HADM ID, and ICU stay ID are unique key values in this case. We focus on extracting details such as prescribed drugs, procedures, and relevant measurements (e.g., blood pressure).

We predict patient mortality along with the following medical diagnoses:

1. Advanced Cancer - Cancers with very high mortality (pancreatic, esophageal, stomach/gastric, biliary, anaplastic).
2. Advanced Heart Disease - Ejection fraction (EF) less than 30%, severe cardiomyopathy, severe aortic stenosis, any mention of heart transplant (considered for, set to receive, denied).
3. Advanced Lung Disease - Pulmonary Function Test (PFT) results of Forced Expiratory Volume (FEV1) less than 50% of normal, or Forced Vital Capacity (FVC) less than 70%. Severe chronic obstructive pulmonary disease (COPD), which may be indicated by Gold Stage III-IV. Severe interstitial lung disease (ILD).
4. Alcohol Abuse - Recent alcohol abuse history which is an active problem at the time of admission, whether it is the primary cause of admission or not.
5. Chronic Neurological Dystrophies - Chronic central nervous system (CNS) or spinal cord diseases, including multiple sclerosis (MS), amyotrophic lateral sclerosis (ALS), muscular dystrophies, myasthenia gravis, Parkinson's Disease, epilepsy, stroke and cerebrovascular accident (CVA) with residual deficits, and various neuromuscular diseases or dystrophies.
6. Chronic Pain Fibromyalgia - Any etiology of chronic pain (including fibromyalgia) requiring long-term opioid/narcotic medication to control.
7. Dementia - Alzheimer's and other forms of dementia are mentioned in the text.
8. Depression - Diagnosis of depression, treatment of depression, presentation to the ICU with symptoms of depression, including acts of self-harm or suicide.
9. Developmental Delay - Includes congenital, genetic, and idiopathic disabilities.
10. Non-Adherence - Temporary or permanent discontinuation of treatment, including pharmaceuticals or appointments, without consulting a physician before doing so. This includes skipping dialysis appointments or leaving the hospital against medical advice. A patient who sees a physician to discuss adverse events associated with a medication may or may not constitute non-adherence, depending on whether the treatment ceased without the physician's consultation.
11. Obesity - Any mention of obesity as a consideration in the healthcare encounter. Abdominal obesity is not sufficient.
12. Other Substance Abuse - Intravenous drug abuse, illicit drug use, accidental overdose of psychoactive or narcotic medications. Remote use of marijuana is not sufficient.
13. Schizophrenia and other Psychiatric Disorders - Psychiatric disorders in DSM-5 classification, including schizophrenia, bipolar, and anxiety disorders. Does not include depression.

We extract the following information for our features leveraging the complicated relational files in the MIMIC-III dataset.

- Patient: Information during the patient's initial hospitalization, such as Subject ID, HADM ID, ICU stay ID, gender, age and obesity.
- Medication: National Drug Code (NDC), Input CV item ID (Liquids injected using the CareVue system), Input MV item ID (Liquids injected using the MetaVision system).
- Procedure: ICD9 billing code assigned to the surgery/procedure undertaken.
- Lab Result: Item ID and Value. The item ID denotes the subject of measurement, e.g., red blood cells. Value for example represents the count of red blood cells for the patient.

The statistics of the dataset are summarized in Table 1.

Table 1: Summary of MIMIC3 Dataset Statistics: A comparison between the dataset preprocessed by our approach and that from the baseline paper. The reported statistics include the average number of unique types of medication, procedures, and lab values connected to each patient, as well as the overall number of unique types across the entire dataset.

Dataset	GraphEHR1	VGNN
Medication	28.9 / 2,042	11.5 / 6,778
Procedures	4.0 / 568	4.5 / 2,006
Lab Values	159.6 / 3,689	62.2 / 2,032
♯ of total patients	1,375	50,391

Label	Number of positives
Mortality	828
Non-Adherence	120
Developmental Delay	29
Advanced Heart Disease	228
Advanced Lung Disease	137
Schizophrenia and Other Psychiatric Disorders	249
Alcohol Abuse	198
Other Substance Abuse	139
Chronic Pain Fibromyalgia	290
Chronic Neurological Dystrophies	324
Advanced Cancer	149
Depression	391
Dementia	104

6 Evaluation Metric

The Area Under the Precision-Recall Curve (AUPRC) is a metric commonly used to evaluate the performance of classification models, particularly in scenarios where imbalanced classes are present. It is calculated based on the precision-recall curve, which illustrates the trade-off between precision and recall for different probability thresholds.

The precision-recall curve is generated by varying the classification threshold and computing precision and recall at each point. Precision is defined as the ratio of true positive predictions to the total number of positive predictions, while recall is the ratio of true positive predictions to the total number of actual positives.

The AUPRC is then computed by integrating the area under the precision-recall curve. Mathematically, the AUPRC is given by:

$$AUPRC = \int_0^1 \text{Precision}(\text{Recall}(t))\, dt \tag{1}$$

Here, Recall(t) represents the recall at a given threshold t, and Precision(Recall(t)) is the precision corresponding to that recall. The AUPRC comprehensively measures a model's ability to balance precision and recall across different classification thresholds.

7 Loss Function

Our study employed the binary cross-entropy loss as the optimization criterion to train various binary classification tasks using the patient node embeddings generated by the Heterogeneous Graph Neural Network (GraphEHR). This loss function is well-suited for binary classification problems, aligning with our objectives of predicting binary outcomes in healthcare tasks such as mortality prediction, ICU readmission prediction, and phenotypic classification.

7.1 Mathematical Description

The binary cross-entropy loss captures the dissimilarity between the predicted probabilities and the actual labels, penalizing the model more severely for misclassifying instances with higher confidence.

$$\mathcal{L}_{\text{Binary Cross Entropy}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2}$$

Let's break down the components of the binary cross-entropy loss:

- y_i: True label for the i-th example (0 or 1)
- $\hat{y}_i$: Predicted probability of the positive class for the i-th example.
- $\log(\hat{y}_i$ Natural logarithm of the predicted probability of the positive class.
- $\log(1 - \hat{y}_i$ Natural logarithm of the predicted probability of the negative class.

The loss is calculated for each example in the sum, and then the average is taken over all examples (N is the number of examples).

7.2 Relation to the Problem Statement

Our binary classification tasks, including mortality, involve predicting critical outcomes based on patient data. The binary cross-entropy loss is a natural choice for such tasks, aligning with the need to effectively penalize and minimize the discrepancies between predicted probabilities and actual outcomes.

8 Baseline Selection

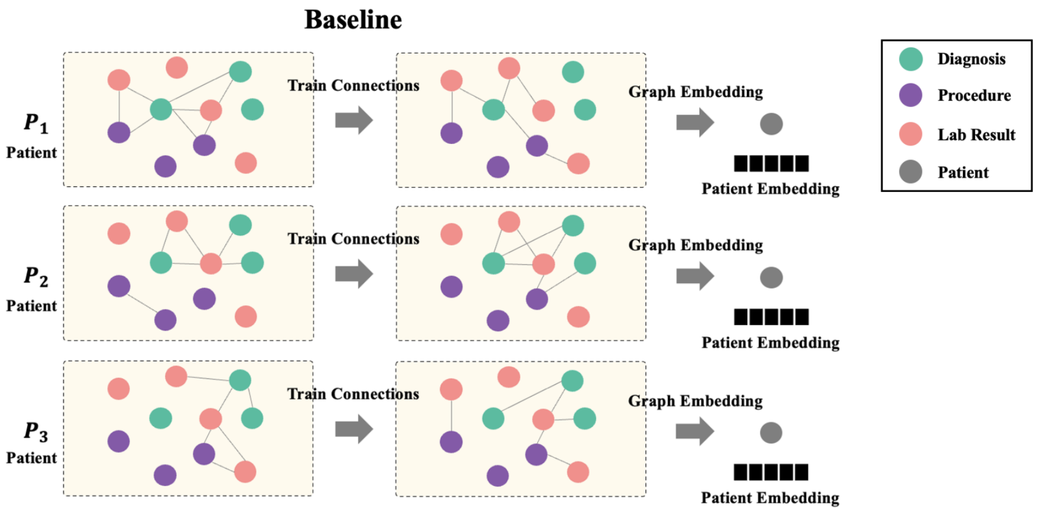

Figure 3: Previous studies mainly concentrated on acquiring patient embeddings based on individual patients. In our proposed approach, we consider node embeddings for each feature and utilize a fusion method to generate embeddings for each patient.

VGNN Basline Model As depicted in Figure 3, the baseline VGNN [29] have constructed graphs for individual patients using features from Electronic Health Records. In detail, each patient is represented by a graph comprising three node types: diagnosis, procedure, and lab results. The quantity of nodes within each category aligns with the respective count of procedures and lab results associated with the patient. The lab results are characterized as continuous values, exemplified by blood pressure levels, and are subsequently categorized into discrete levels: high, medium, or low. It

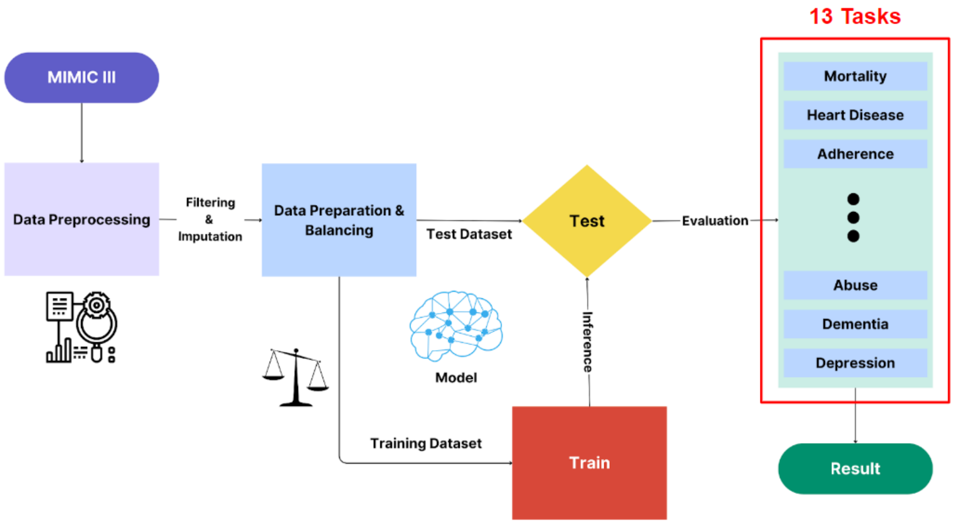

Figure 4: This flow diagram summarizes our implementation from efficiently transforming MIMIC-III EHR data through pre-processing and addressing class imbalance to training a Heterogeneous Graph Neural Network. Our adaptive approach attempts multiple predictive tasks, empowering accurate insights into patient health.

integrates an encoder and decoder with a variational regularization. The encoder processes a complete graph with interconnected nodes, employing a self-attention mechanism to update node embeddings. The variational regularization is trained to establish a prior distribution for the initial node features of the decoder, utilizing the output node embeddings from the encoder. Subsequently, latent variables sampled from the prior distribution serve as inputs to the decoder. An additional node is introduced to the graph to capture the patient's embedding, establishing connections with all existing nodes in the decoder. The decoder employs the patient node as a query, while other nodes function as key and value inputs, ultimately utilizing the patient's embedding for the corresponding patient's label prediction. However, it faces challenges in reflecting the importance of each feature as the number of features increases. Creating distinct graphs for each patient as their embedding makes it less conducive to analyzing commonalities and differences among patients. Additionally, it constructs a homogeneous graph despite having different types of nodes in Electronic Health Records features.

In response, we propose a novel heterogeneous graph embedding approach by performing node embedding for each feature, capturing relationships between each feature and all patients. Specifically, we distinguish features in Electronic Health Records as different nodes and discern their relationships with all patients ass illustrated in Figure 2 We leverage this representation to perform various phenotype prediction tasks, validating the effectiveness of our new graph-based representation learning in accurately capturing patient characteristics.

Other Baseline Models The most competitive baseline we are testing against is VGNN [29], where each patient is represented by a graph comprising three node types: diagnosis, procedure, and lab results. Based on our literature review, the following other methods can also be used as baseline models to statistically compare the performance of our methodology. To compare the performance of our model on different prediction tasks (mortality, phenotypes, etc.), we will be individually training them for each prediction task.

- Logistic Regression: Logistic Regression with balanced class weights and specified regularization parameters for binary classification tasks.
- Random Forest [1]: Random Forest classifier with balanced class weights and a specified number of estimators for ensemble learning in classification tasks. It leverages decision trees to make predictions.

- Gradient Boosting: Gradient Boosting classifier with a specified number of estimators for ensemble learning in classification tasks. It sequentially builds weak learners to boost overall model performance.
- SVM (Support Vector Machine with Standard Scaler): Support Vector Machine classifier with balanced class weights, a specified regularization parameter (C), and feature scaling using Standard Scaler. It aims to find a hyperplane that best separates classes.
- KNN (K-Nearest Neighbors with Standard Scaler): K-Nearest Neighbors classifier with a specified number of neighbors and feature scaling using Standard Scaler. It classifies data points based on the majority class of their k-nearest neighbors.
- MLP1 to MLP5 (Multilayer Perceptrons with Standard Scaler) [25]: Multilayer Perceptron classifiers with varying hidden layer sizes, logistic activation, regularization parameters, and feature scaling using Standard Scaler. These models are known for their ability to capture complex relationships in data.

We conducted experimental analysis on selected baselines to assess their performance against our GNN model. Additionally, we utilized the GNN from the reference paper, [29], as a benchmark for evaluating our GNN model. Our and baseline models performed 13 prediction tasks on the same benchmark.

9 Baseline Implementation

We trained the baseline model VGNN [29] based on the original code, but some adjustments were necessary to apply to our own dataset. We trained VGNN for 50 epochs, retaining the initially proposed optimal hyperparameters except for the embedding size dimension. Due to memory constraints, we opted for a reduced embedding dimension of 128 instead of the original 256. For other baseline models such as Logistic Regression, Multi-Layer Perceptrons etc. we baked in SMOTE over-sampling of the minority class and under-sampling of the majority class to do away with the class imbalance issue. We also added regularization into the baseline models. The five Multi-Layer Perceptrons architectures corresponded to variable depths and nodes per layer. MLP1 was a single hidden-layer neural network with 100 hidden neurons. MLP2 had two layers each with 50 neurons, while MLP3 had two layers with 100 and 50 neurons each. Both MLP4 (3 layers with 200, 100, and 50 neurons) and MLP5 (4 layers with 300, 200, 100, and 50 neurons) represented the historical sense of deep neural networks with more than 3 hidden layers.

10 Implemented Extensions & Experiments

GraphEHR1 For GraphEHR1, we trained over 50 epochs, mirroring the baseline model's training duration. Utilizing a configuration with 8 heads, an embedding size of 256, and an exploration of 9 distinct learning rates (ranging from 0.0005 to 0.01), coupled with 3 different layer depths (1, 2, 3). It's noteworthy to highlight that despite additional hyperparameters for our model, the computational cost for the baseline was notably high.

GraphEHR2 For the second approach, GraphEHR2, as discussed, we included patient-patient edges as well and included lab values as a separate node type. The learning rate was kept at 0.001 with a weight decay of $5e^{-3}$ using an AdamW optimizer. All the models were trained for 200 epochs, and the model with the best validation AUPRC was chosen. The hidden channel dimension for the Graph Transformer was kept at 64 with 4 attention heads and 2 Transformer Layers.

The code we executed can be found at `https://github.com/kushagragarwal2443/GraphEHR`.

11 Results and Discussion

The detailed outcomes are presented in Table 2. Our proposed models, GraphEHR(1) and GraphEHR(2), manifest superior performance across nearly all evaluated tasks compared to baseline models. In tasks such as predicting advanced heart disease, schizophrenia, other psychiatric disorders, and alcohol abuse, VGNN demonstrates commendable performance relative to GraphEHR. However, it is crucial to note that the observed performance gap is insignificant. Notably, as evident in the

Table 2: Model evaluation of mortality prediction in our dataset using precision-recall curves (99% confidence interval)

Algorithm	LR	RF	GB	SVM	KNN	MLP1	MLP2	MLP3	MLP4	MLP5	VGNN	GraphEHR(1)	GraphEHR(2)
Mortality	0.68	0.65	0.68	0.72	0.61	0.71	0.69	0.73	0.69	0.67	0.76	0.81	**0.86**
Non-Adherence	0.08	0.07	0.09	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.23	0.30	**0.31**
Developmental Delay	0.33	0.01	0.10	0.01	0.01	0.25	0.33	0.25	0.20	0.01	0.23	**0.34**	0.29
Advanced Heart Disease	0.12	0.10	0.12	0.10	0.11	0.15	0.15	0.13	0.11	0.19	**0.56**	0.40	0.53
Advanced Lung Disease	0.10	0.12	0.10	0.09	0.10	0.13	0.11	0.14	0.11	0.13	0.16	0.16	**0.39**
Schizophrenia and Other Psychiatric Disorders	0.17	0.13	0.17	0.13	0.11	0.15	0.14	0.14	0.14	0.14	**0.46**	0.32	0.44
Alcohol Abuse	0.35	0.24	0.25	0.24	0.15	0.33	0.29	0.31	0.30	0.32	**0.52**	0.43	0.47
Other Substance Abuse	0.11	0.06	0.08	0.06	0.07	0.08	0.11	0.11	0.09	0.07	0.36	0.25	**0.42**
Chronic Pain Fibromyalgia	0.17	0.15	0.15	0.16	0.16	0.18	0.19	0.20	0.19	0.19	0.21	0.30	**0.44**
Chronic Neurological Dystrophies	0.23	0.24	0.25	0.24	0.25	0.23	0.27	0.24	0.27	0.28	0.29	0.25	**0.48**
Advanced Cancer	0.08	0.08	0.08	0.08	0.09	0.09	0.09	0.11	0.10	0.09	0.15	0.12	**0.30**
Depression	0.39	0.38	0.36	0.36	0.36	0.41	0.39	0.40	0.38	0.37	0.32	0.41	**0.54**
Dementia	0.10	0.11	0.26	0.09	0.09	0.29	0.22	0.27	0.29	0.15	0.08	0.16	**0.27**

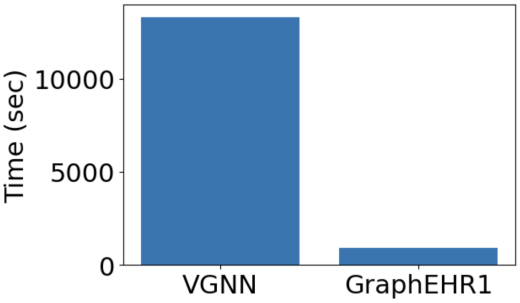

Figure 5: This figure demonstrates that GraphEHR significantly outperforms VGNN regarding computational speed in mortality prediction tasks.

results, GraphEHR surpasses VGNN by more than two times in predicting advanced lung disease or dementia. Moreover, as shown in Figure 5, the running time of GraphEHR is much smaller than that of VGNN. Consequently, our proposed method demonstrates enhanced efficacy, particularly in handling the complexities of the MIMIC-III dataset.

Furthermore, it is noteworthy that the overall Area Under the Precision-Recall Curve (AUPRC) tends to be lower than the mortality prediction task. This discrepancy arises from the inherent class imbalance in other tasks, as illustrated in Table 1, reducing positive instances. We observe that baseline models exhibit heightened vulnerability in the presence of class imbalance. Specifically, the VGNN, reliant on learning connections between nodes, appears to be particularly sensitive to the quality of the training dataset, especially in scenarios where label imbalance exists. This vulnerability is evident in instances such as the dementia prediction task, where VGNN yields a notably low AUPRC of 0.08, underscoring the model's potential challenges in capturing pertinent connections under these circumstances. Figure 6 compares the models across the 13 predictive tasks.

Lastly, the re-implemented baseline yields higher results than those documented in the original paper for the mortality prediction task. This disparity can likely be attributed to our dataset's reduced number of patients, resulting in a smaller test dataset. Consequently, obtaining a higher Area Under the Precision-Recall Curve (AUPRC) is reasonable.

A very important design consideration we made was to include two GraphEHR architectures to motivate the transition from the baseline model, VGNN to what we hypothesized the best model would be, GraphEHR2. If one carefully thinks, GraphEHR1 in theory should work better than VGNN but not with a huge difference as it is using the same information from the EHR in a similar way. However, the only difference was that it kept all the patients in a single heterogeneous graph (node embeddings) instead of creating separate graphs for each patient (graph embeddings in VGNN). This allowed some degree of information sharing between patients via connections through other node types. However, we knew that this would not be able to yield the most superior results, as we are still missing out on information sharing between readmissions. Therefore, after adding the edges between identical patient readmissions, and keep lab values a separate node type (design choice), we observed much superior results as presented above.

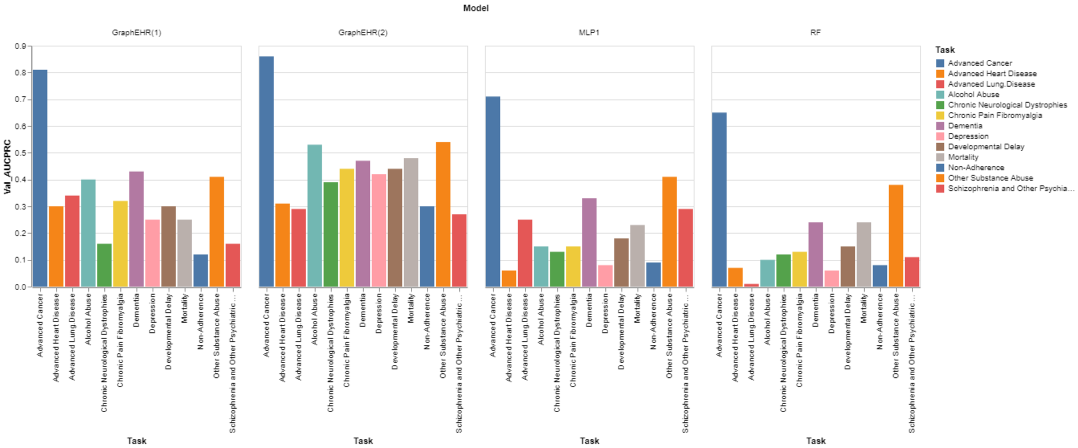

Figure 6: From this figure we can compare the performance of the baselines with our proposed GraphEHR models. We can see that GraphEHR1 and GraphEHR2 significantly outperform the baselines across the tasks.

As the table shows, simplistic models such as Logistic Regression, Support Vector Machines, and even tabular Multi-Layer Perceptrons fail to capture the complexities present in EHR data. This necessitates the adoption of more expressive models. Additionally, EHR data often exhibits inconsistent variable lengths across different features, making it essential to employ an architecture capable of handling such variability. The only architecture suitable for this modality is a graph. To harness the versatility of graphs and enhance expressiveness, we naturally gravitated towards heterogeneous graphs, which can accommodate multiple node types and edge relations.

From the aforementioned results, it is apparent that GraphEHR2 outperformed all other models in 9 out of 13 tasks, while GraphEHR1 excelled in 1 task, and VGNN in 3. There remains considerable room for improvement in this domain, especially with the inclusion of additional modalities such as CT-Scans and ECGs. This would unveil the rich information concealed in the Pandora's box of EHR data, potentially leading to significant advancements in patient outcomes and making a meaningful impact.

12 Future Works

As part of our future work, we intend to enhance the interpretability of learned representations to facilitate understanding by healthcare professionals. We aim to integrate interpretability techniques, such as attention mechanisms, to highlight influential nodes or edges in the heterogeneous graph. We hope this will provide insights into why certain relationships contribute to predictions. This will increase trust in the model's decision-making process, making it more accessible for healthcare practitioners. Furthermore, upon EDA, we discovered that ICD-9 codes [14] are not medically accurate, so we had to use a subset of data for which the actual medical diagnosis was available. However, this subset was not very huge, hence as future work, we will try to incorporate other datasets except MIMIC like eICU.

13 Conclusion

Our project attempted to develop a transformative approach in healthcare analytics using EHR data. We're harnessing the power of Graph Neural Networks (GNN) to unravel the intricate web within Electronic Health Records. Our Heterogeneous Graph Network ensures a robust understanding of complex relationships, leading to adaptive performance enhancements across diverse predictive tasks. The profound impact leads to enhanced diagnosis, improved treatment, and optimized patient care. Looking ahead, our use of Heterogeneous GNN opens doors to adaptive multitasking, promising accurate and insightful applications. We hope that our work benefits both medical professionals and patients, empowering decision-making and elevating the quality of healthcare outcomes.

14 Division of Work

- Kushagra Agarwal: Problem formulation. Worked on data pre-processing. Implented and trained other baseline models such as MLPs, SVM, Random Forest. Implemented and trained GraphEHR2.
- Minyoung Choe: Problem formulation, worked on VGNN baseline model implementation and training, GraphEHR1 implementation and training.
- Nivedhitha Dhanasekaran: Literature Review for problem formulation, data pre-processing & optimization, baseline model hyperparameter optimization implementation & training, and documentation.
- Juho Jung: Worked on data pre-processing, baseline model comparative analysis, and video preparation.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[3] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[4] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 432–440. SIAM, 2016.

[5] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[6] Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *Advances in neural information processing systems*, 31, 2018.

[7] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 606–613, 2020.

[8] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[9] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[10] Alexander Hoerbst and Elske Ammenwerth. Electronic health records. *Methods of information in medicine*, 49(04):320–336, 2010.

[11] Md. Ekramul Hossain, Arif Khan, Mohammad Ali Moni, and Shahadat Uddin. Use of electronic health data for disease prediction: A comprehensive literature review. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2):745–758, 2021. doi: 10.1109/TCBB.2019.2937862.

[12] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pages 2704–2710, 2020.

[13] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177: 141–153, 2019.

[14] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[15] Deepak A Kaji, John R Zech, Jun S Kim, Samuel K Cho, Neha S Dangayach, Anthony B Costa, and Eric K Oermann. An attention based deep learning model of clinical events in the intensive care unit. *PloS one*, 14(2):e0211057, 2019.

[16] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[17] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[18] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[19] Sheng-Chieh Lu, Cai Xu, Chandler H Nguyen, Yimin Geng, André Pfob, and Chris Sidey-Gibbons. Machine learning–based short-term mortality prediction models for patients with cancer using electronic health record data: systematic review and critical appraisal. *JMIR medical informatics*, 10(3):e33182, 2022.

[20] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in neural information processing systems*, 32, 2019.

[21] Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Machine learning for healthcare conference*, pages 73–100. PMLR, 2016.

[22] Matthieu Scherpf, Felix Gräßer, Hagen Malberg, and Sebastian Zaunseder. Predicting sepsis with a recurrent neural network using the mimic iii database. *Computers in biology and medicine*, 113:103395, 2019.

[23] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[24] Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[25] Navdeep Tangri, David Ansell, and David Naimark. Predicting technique survival in peritoneal dialysis patients: comparing artificial neural networks and logistic regression. *Nephrology Dialysis Transplantation*, 23(9):2972–2981, 2008.

[26] Claire Thornton. 795,000 americans a year die or are permanently disabled after being misdiagnosed, Jul 2023. URL `https://www.usatoday.com/story/news/health/2023/07/18/medical-misdiagnosis-killing-disabling-americans/70423573007/`.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[28] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[29] Weicheng Zhu and Narges Razavian. Variationally regularized graph-based representation learning for electronic health records. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 1–13, 2021.