Artificial Intelligence In Medicine 135 (2023) 102439

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

An integrated LSTM-HeteroRGNN model for interpretable opioid overdose risk prediction

Xinyu Dong [a], Rachel Wong [b], Weimin Lyu [a], Kayley Abell-Hart [b], Jianyuan Deng [b], Yinan Liu [a], Janos G. Hajagos [b], Richard N. Rosenthal [c], Chao Chen [a,b,*], Fusheng Wang [a,b,**]

[a] Department of Computer Science, Stony Brook University, Stony Brook, NY, United States of America
[b] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States of America
[c] Department of Psychiatry, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, United States of America

ARTICLE INFO

Keywords:
Opioid overdose
Opioid poisoning
Deep learning
Clinical decision support
Electronic health records
Long short-term memory
Graph neural network

ABSTRACT

Opioid overdose (OD) has become a leading cause of accidental death in the United States, and overdose deaths reached a record high during the COVID-19 pandemic. Combating the opioid crisis requires targeting high-need populations by identifying individuals at risk of OD. While deep learning emerges as a powerful method for building predictive models using large scale electronic health records (EHR), it is challenged by the complex intrinsic relationships among EHR data. Further, its utility is limited by the lack of clinically meaningful explainability, which is necessary for making informed clinical or policy decisions using such models. In this paper, we present LIGHTED, an integrated deep learning model combining long short term memory (LSTM) and graph neural networks (GNN) to predict patients' OD risk. The LIGHTED model can incorporate the temporal effects of disease progression and the knowledge learned from interactions among clinical features. We evaluated the model using Cerner's Health Facts database with over 5 million patients. Our experiments demonstrated that the model outperforms traditional machine learning methods and other deep learning models. We also proposed a novel interpretability method by exploiting embeddings provided by GNNs to cluster patients and EHR features respectively, and conducted qualitative feature cluster analysis for clinical interpretations. Our study shows that LIGHTED can take advantage of longitudinal EHR data and the intrinsic graph structure of EHRs among patients to provide effective and interpretable OD risk predictions that may potentially improve clinical decision support.

1. Introduction

The United States is experiencing an opioid crisis, with an estimated 10 million people aged 12 or older misusing opioids in 2019 [1], and 130 deaths a day from opioid overdose (OD) [2]. The estimated economic burden of the opioid epidemic is approximately 78.5 billion dollars per year [3]. Early intervention to reduce the risk of OD can help to decrease opioid-related mortality, and building a predictive risk model for OD can help to inform interventions. Large-scale electronic health records (EHRs) provide enormous amounts of data that can be used to build data-driven models. Such models can identify patients at high risk of OD and reveal critical explanatory knowledge for the prediction.

In recent years, deep learning methods have gained popularity in building predictive models to assist clinical decision support. Sequential models such as the recurrent neural network (RNN) have been widely applied to diseases such as Parkinson, Alzheimer's, and heart disease [4–6] for their ability to incorporate the temporal effects of disease development. Advanced RNN models like long short-term memory (LSTM) models have been shown to have high performance in opioid related disease prediction [7–9].

In addition to the temporal dynamics of EHR data, complex interactions and extreme sparsity of features are also major challenges for developing predictive models with EHR data. Graph neural networks (GNNs) are frequently used in building predictive models on data that can be described by graphs [10], and hold the potential to address the

* Correspondence to: C. Chen, Department of Biomedical Informatics, Stony Brook University, 2313C Computer Science, Stony Brook, NY 11794-8330, United States of America.
** Correspondence to: F. Wang, Department of Biomedical Informatics, Department of Computer Science, Stony Brook University, 2313D Computer Science, Stony Brook, NY 11794-8330, United States of America.
E-mail addresses: chao.chen.1@stonybrook.edu (C. Chen), fusheng.wang@stonybrook.edu (F. Wang).

https://doi.org/10.1016/j.artmed.2022.102439
Received 11 January 2022; Received in revised form 27 September 2022; Accepted 28 October 2022
Available online 3 November 2022

above challenges. Traditional deep neural networks such as convolutional neural networks (CNNs) and RNNs assume an underlying domain with a regular structure, but for graph-structured data, the local neighborhood of each node does not have a fixed connectivity. Various types of GNNs have been proposed to address this issue, such as graph convolutional networks (GCN) and graph attention transformers (GAT), resulting in powerful predictive models [11–13]. During the learning phase of GNNs, the feature representation of each node is iteratively updated based on information from itself and from neighbors. The information from a neighbor, called the message, is a linear transformation of the neighbor's representation in its previous iteration. Furthermore, heterogeneous and relational GNN models can set different types to nodes and edges and assign different parameters to them, then learn the variation with different parameters depending on the type of edge within the graph. With these more complex structures, rich information can be encoded in graphs, which achieve better prediction performance [14–16]. We envision that GNN can be useful in clinical practice for two reasons. The complex relations between patients sharing similar health conditions and diagnoses can be represented by a graph, which can be well represented by the graph structure of a GNN model. Moreover, GNN is effective in tackling the sparsity problem in EHR data. In the traditional tabular method to represent EHR data, the input data are extremely sparse as an indivieual patient is mostly associated with a a limited set of clinical codes or tests while a large number of features are available in the EHR database. GNN can help to address the sparsity issue since patients are not connected to features that are absent in their encounters.

Because of the difficulty in interpreting complex deep learning models and the demand to find clinical meanings from the models, along with the development of more advanced deep learning models, many efforts have been made to interpret the models and give clinically meaningful explanations. Some methods attempt to give explanations indirectly; for example, model-agnostic methods use an interpretable model to simulate deep learning models and use the interpretable model as a proxy to provide explanations [17]. Other interpretation methods measure feature importance based on the changes in prediction performance after manipulating distributions of input feature values, such as the Local Surrogate model (LIME) [18] and the permutation importance method [19].

In this work, we propose a **l**ong**i**tudinal and **g**rap**h** in**te**grate**d** (LIGHTED) prediction model using electronic health records that combines LSTM and GNN for predicting opioid overdose risk in patients who have been prescribed opioid medications. LIGHTED takes advantage of not only the longitudinal history of patients' EHRs, but also the knowledge learned from relationships among patients, encounters, and different types of features. Our experimental results demonstrate that our sequential deep learning model provides superior performance, with a F1 score higher than traditional machine learning methods as well as other state-of-the-art deep learning models. We also propose a novel interpretability approach in which we group features and patients respectively using GNN representations, generating less redundant and more semantically meaningful feature clusters for clinical interpretations and informing clinical decision support.

2. Dataset description

2.1. Data source

We chose Cerner's Health Facts database [20] as our data source to build the dataset. It is one of the largest EHR databases, with deidentified data from >65 million patients and over 600 different healthcare facilities in the United States. It includes records of patient demographics, encounters, diagnoses, prescriptions, procedures, laboratory tests, and medical billing information.

2.2. Study population

In this study, we use EHR data from January 1, 2008 to December 31, 2017. For the patient cohort, we extracted all patients who were prescribed medications containing active opioid ingredients. We used the Anatomical Therapeutic Chemical (ATC) level 3 code 'N02A' and category description 'opioid' to retrieve all relevant active ingredients from DrugBank 5.1.4 [21]. To define opioid poisoning, we used a group of ICD-9 and ICD-10 codes identified by Moore [22] to select patients with opioid overdose.

The distributions of the first opioid medication exposure between non-OD and OD patients are not identical; compared to OD patients, non-OD patients have a much higher proportion of patients younger than 16 years and a much lower proportion of patients who are older than 66 years [23,24]. To make the distribution more consistent and also to prevent bias toward age, we filtered the patients to include those between 16 and 66 [8].

Patients with a cancer diagnosis were excluded. They often have acute and varying symptoms of pain due to disease or treatment which requires opioid therapy for pain control. As medical practitioners and public health leaders have largely focused on guidelines for patients with chronic non-cancer pain, we excluded patients with cancer diagnoses from our model [25]. We identified cancer diagnoses based on cancer related ICD-9 [26] and ICD-10 codes [27] (Supplementary Table 2). The flowchart for patient selection is shown in Fig. 1.

2.3. Feature selection

We extracted the following five categories of clinical features for prediction: diagnosis codes, procedure codes, lab tests, medications, clinical events, and demographic information. The following paragraphs detail the features in each category and how they were processed. Our data preprocessing follows the methods used in our prior work [7,9].

Diagnosis codes specify patients' diseases and symptoms, which contain critical information for predicting future clinical events such as OD. There are two versions of ICD codes, ICD-9 and ICD-10, used in Health Facts to record diagnoses. We converted ICD-9 codes to ICD-10 when they indicate the same clinical meaning to avoid dispersion of predictability [28]. Then, in order to reduce granularity, we extracted the first 3 digits of each code as features.

Medications are recorded by National Drug Code (NDC) codes in Health Facts. The NDC code is a universal product identifier for human drugs specific to the individual label. Instead of manufacturer or packager level information, a clinically meaningful representation is more useful to predictive models, so we converted all NDC codes to

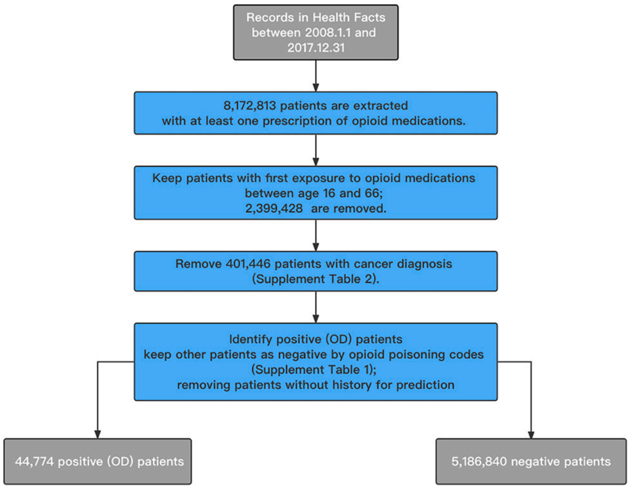
Records in Health Facts between 2008.1.1 and 2017.12.31
8,172,813 patients are extracted with at least one prescription of opioid medications.
Keep patients with first exposure to opioid medications between age 16 and 66; 2,399,428 are removed.
Remove 401,446 patients with cancer diagnosis (Supplement Table 2).
Identify positive (OD) patients keep other patients as negative by opioid poisoning codes (Supplement Table 1); removing patients without history for prediction
44,774 positive (OD) patients
5,186,840 negative patients

Fig. 1. Flowchart of selecting patients.

Anatomical Therapeutic Chemical (ATC) codes. An ATC code is a unique code indicating the active ingredients according to the system/organ they target and how they work. Moreover, ATC level 3 codes were chosen [29] to represent all medications, because level 3 is the most detailed therapeutic/pharmacological class, whereas higher levels (i.e., 4,5) are for chemical classes. For each medication, the total medication quantity prescribed to each patient was calculated as a feature for each medication. In addition, morphine milligram equivalents (MME) were specifically calculated as an aggregate feature. MME is a value assigned to opioids representing their relative potencies compared to morphine [30,31], which is an informative standardized indicator of a patient's overall opioid intake. We summed all MMEs for each encounter as the value for that feature.

Lab Tests record a sample of a patient's blood, urine, other bodily fluid or body tissue, and analyze the sample to gain information about the patient's health. Common lab tests include complete blood count, prothrombin time, hemoglobin, etc. For each lab test, Health Facts not only provides a numeric value result, but also a standardized interpretation of the value indicating whether it is high, low, or normal. We calculated the portion of high and low values that each patient received for each test, as well as the total number of tests that the patients received.

Clinical events are other related personal health situations that are not formally classified into medical codes, like height, weight, and BMI. Some clinical events are highly relevant to opioid use, such as pain score indicating the pain level, and substance use history, which includes tobacco and alcohol.

Demographic information was extracted as features, including age, gender, and race/ethnicity.

To prevent overfitting, features of low co-occurrence with an opioid overdose diagnosis were removed, which we defined as <1 % prevalence in prior records of all opioid overdose patients. For this filtering threshold of 1 %, we performed an experiment to test the prediction performance with a threshold from 0.5 % to 5 % and also with a fixed number of features from 100 to 2000 in previous work [7–9]. Based on the LSTM model's performance, we had the best result with a threshold of 1 %. To keep the work consistent and comparable, we continue with this threshold of 1 %. After the removal process, 1185 features remained, which included 414 diagnosis codes features, 394 laboratory test features, 3 demographic features, 227 clinical event features, and 147 medication features (summarized in Table 1).

2.4. Encounter selection for input construction

For input feature matrix construction, we first identified the encounter where the prediction target, or diagnosis of opioid poisoning, first occurred. For positive cases, the prediction encounter was the first encounter with a diagnosis of opioid poisoning. For negative cases, the last encounter was used as the target encounter. We then selected the last 5 encounters that occurred at least 14 days and at most 12 months before the target encounter to build the feature vectors (Fig. 2). If a patient had fewer than 5 encounters for the prediction, we repeated the last available encounter to fill the gap. For instance, if a patient had only 3 encounters prior to the prediction encounter, then the feature matrix for the patient would be composed of one feature vector for the first encounter, one feature vector for the second encounter, and three feature vectors replicated from the third encounter [7–9]. We evaluated different numbers of encounters, and 5 was the optimal number with the best performance. The 12 months' constraint is a common standard in previous opioid overdose studies, which is also employed in insurance claims and Veterans Health Administration areas [32–35]. The 14 days' constraint is employed to exclude encounters with a close time frame to the date of the target.

Table 1
Summary of features.

Datasets	Category	# of Features	Description
Health Facts	Diagnosis	414	First 3 digits of ICD-10 (ICD-9 codes were converted to ICD-10)
	Laboratory Test Result	394	Number of high, low and normal values for each test
	Demographics	3	Gender, Age, Race/Ethnicity
	Clinical Events	227	Other related personal health situations that are not formally classified into medical codes, e.g., height, weight, BMI, smoking history and other substance use history. Continuous variables like height, weight, BMI are recorded by their numeric value.
	Medication	147	Total quantity prescribed for each medication

3. Methodology

3.1. Input feature matrix construction

Our data input is composed of two parts; we denote them as 1) raw feature matrix and 2) graph embeddings. The raw feature matrix is built by following the procedure in the previous section to record the existence or value of mentioned clinical features. Graph embeddings are obtained by the GNN model described next to represent the interactions of features and patients. Although we split them into two parts, in actuality, the GNN model learns graph embeddings from the graph built through extracting feature relations from the raw feature input. The design of how we build each input feature matrix and feed them to the models is shown in Fig. 3.

3.2. Heterogeneous relational graph model construction

Heterogeneous GNNs are a variant of GNN models that contain different types of nodes and edges and are capable of capturing the characteristics of different node and edge types [14,15]. We incorporated 3 different types of nodes: patient, encounter and feature, as shown in Fig. 4.

Each patient node represents a patient and each encounter node represents one encounter for one patient. For each pair of node types, we defined an edge type, such as a patient-encounter edge between patient and encounter nodes, or an encounter-feature edge between encounter and feature nodes.

To initiate the feature node vectors, we used a one-hot encoding vector. The i-th entry with the value of 1 indicates a specific lab test was performed or a specific diagnosis was given (as shown in Fig. 5), and the value of 0 indicates it was not present. For example, to represent the 414 different diagnosis features, the vector for the diagnosis node had a length of 414 with value 1 or 0 indicating the presence or absence of a diagnosis respectively. Initial input vectors to nodes were then projected into a shared latent space with a trainable weight matrix W according to different feature types. Inspired by relational GCN models, we defined the following propagation model for calculating the forward-pass update of nodes:

$$h_i^{l+1} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}W_r^l h_j^l + W_0^l h_i^l\right)$$

h_i^{l+1} is the hidden state of the node i in layer $l+1$, while h_i^l is for layer l. N_i^r denotes the set of neighbor indices of node i under relation r. For each edge type, there is a weight vector W_r^r indicating the linear transformation of that edge type. To ensure that the representation of a node at layer $l+1$ can also be informed by the corresponding representation at layer l, we added weight W_0^l to the representation of layer l. $c_{i,\ r}$ is a problem-specific normalization constant that can either be learned or

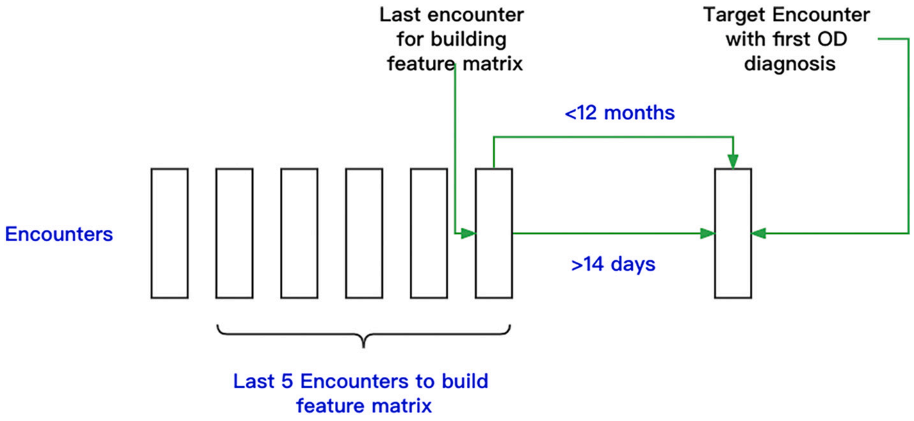
Last encounter for building feature matrix
Target Encounter with first OD diagnosis
<12 months
Encounters
>14 days
Last 5 Encounters to build feature matrix

Fig. 2. Encounters for prediction and building feature matrix.

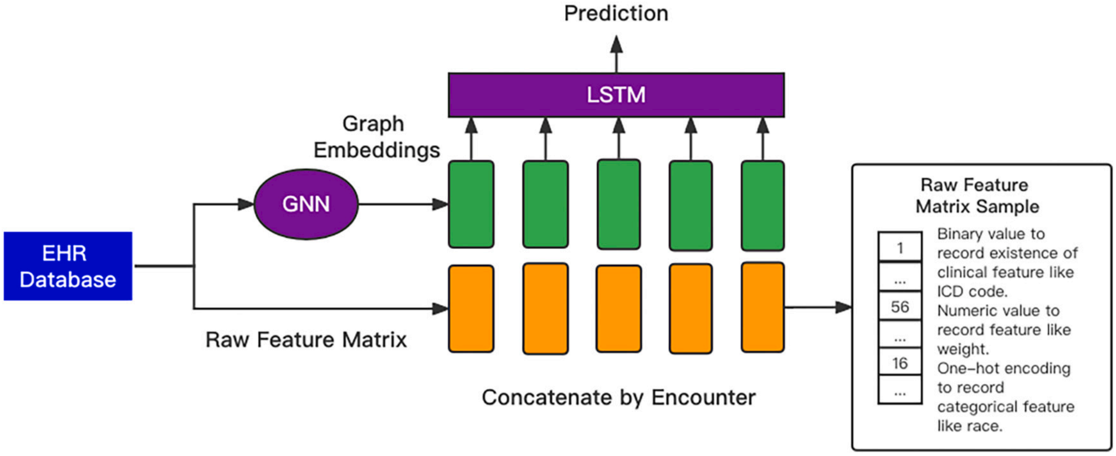
Prediction
LSTM
Graph Embeddings
GNN
EHR Database
Raw Feature Matrix
Concatenate by Encounter
Raw Feature Matrix Sample
1
...
56
...
16
...
Binary value to record existence of clinical feature like ICD code.
Numeric value to record feature like weight.
One-hot encoding to record categorical feature like race.

Fig. 3. Structure of input feature matrix and model.

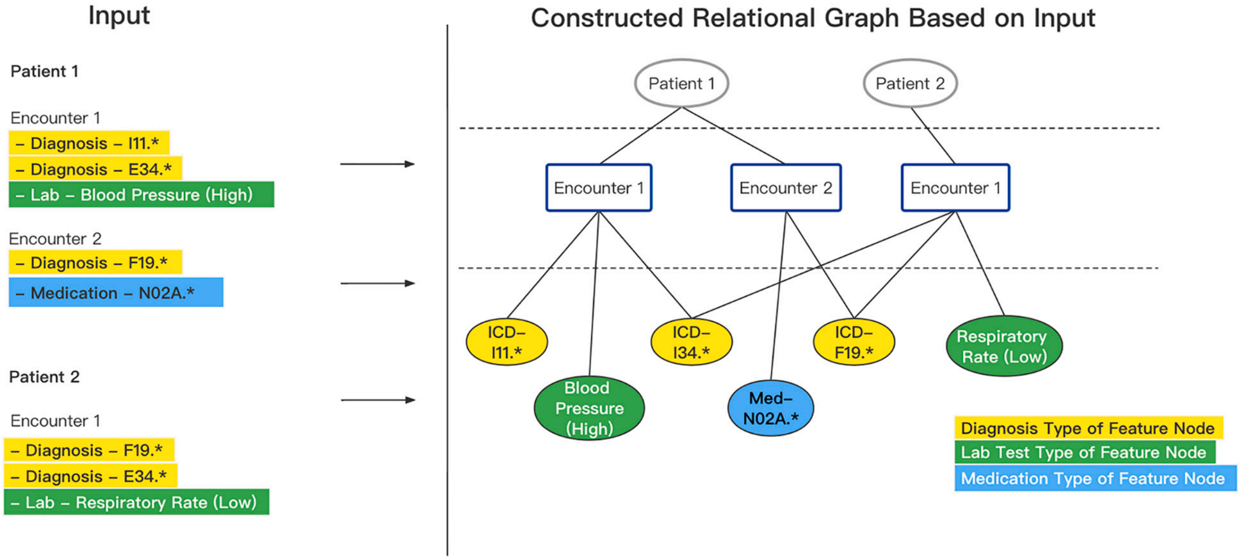

Fig. 4. Structure of heterogeneous relational graph.

chosen in advance, for which we use $c_{i,r} = |N_i^r|$. During the training phase, we obtained the weights for different types of edges. During the test phase, we reconstructed the graph based on the test set with different edge connections and the same edge weights.

3.3. Integration with LSTM model

We integrated the embeddings learned from the Heterogeneous graph with the LSTM model to form our proposed HeteroRGCN+LSTM

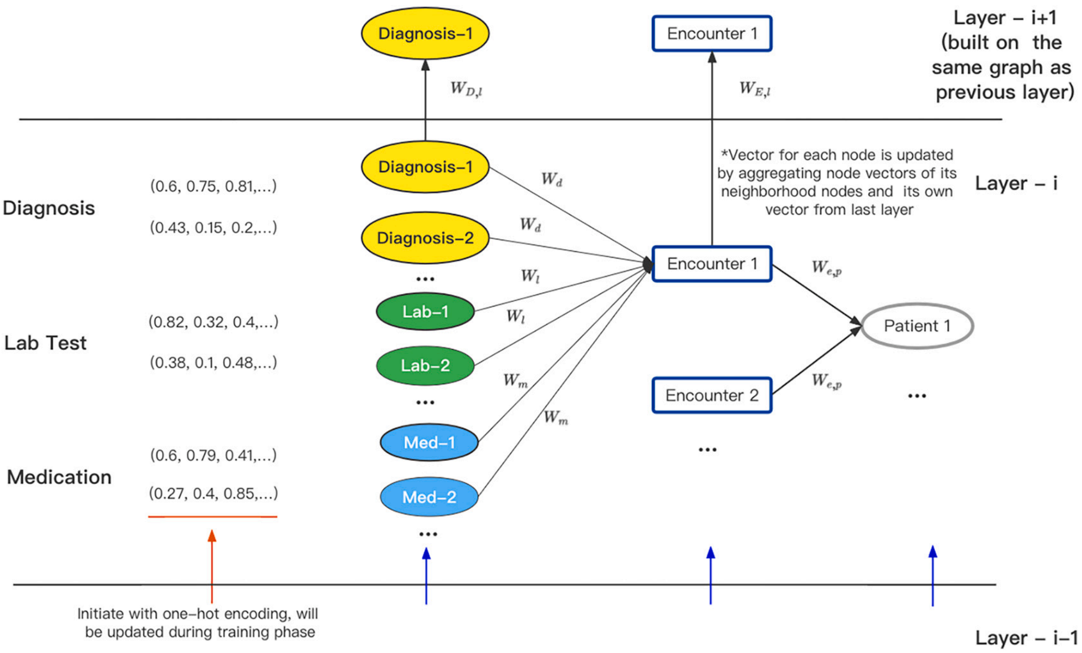

Fig. 5. An example visual representation of features in a heterogeneous graph. The vector for each node is updated by aggregating node vectors of its neighborhood nodes and its own vector from the last layer. The feature node vector will be initiated by one-hot encoding.

model, which we entitled “LIGHTED”. Long Short-Term Memory (LSTM) networks are a version of RNNs. LSTM has two main advantages over common RNNs; LSTM has better memory of knowledge learned from past inputs, and LSTM can solve the vanishing gradient problem faced by RNNs [36]. The embedding of patient encounter nodes was concatenated with original feature inputs, together forming the input features for the LSTM model. The structure of the proposed LIGHTED model is shown in Fig. 6. The parameters of the LSTM and graph were trained simultaneously. We implemented the integrated network with two layers of HeteroRGCN with 100-dimensional embeddings and two layers of LSTM with 64 hidden units. The whole model was trained with binary cross-entropy loss function and Adam optimizer. These parameters were chosen for their performance in preliminary experiments.

3.4. Interpretability

Unlike inferential models, deep learning based predictive modeling can result in a “black box” with predictions difficult to interpret [37]. A common approach is to rank features based on the importance of the predictions, through methods such as permutation importance [19] and Shapley values [44]. However, feature importance ranking may not be informative enough for clinical use. First, some top features may not be available in the EHR for a specific patient due to the sparse nature of EHR data. Second, values of some features may have multiple clinical meanings; for example, a higher weight may result from a variety of different health conditions. Lastly, the large number of features may be intimidating or infeasible for clinicians to interpret, considering that more than a thousand EHR features are used in the predictions.

To address these challenges, we first provide a global importance ranking for all features, based on permutation importance, Shapley value, and model-agnostic methods. Then, we apply a clustering algorithm to group features based on graph embeddings extracted from the LIGHTED model. Graph embedding is an approach to transform features and their graphic information into a vector space, and use those transformed vectors as representations of the features. It is commonly achieved by the inner output from a GNN model [55]. When building a

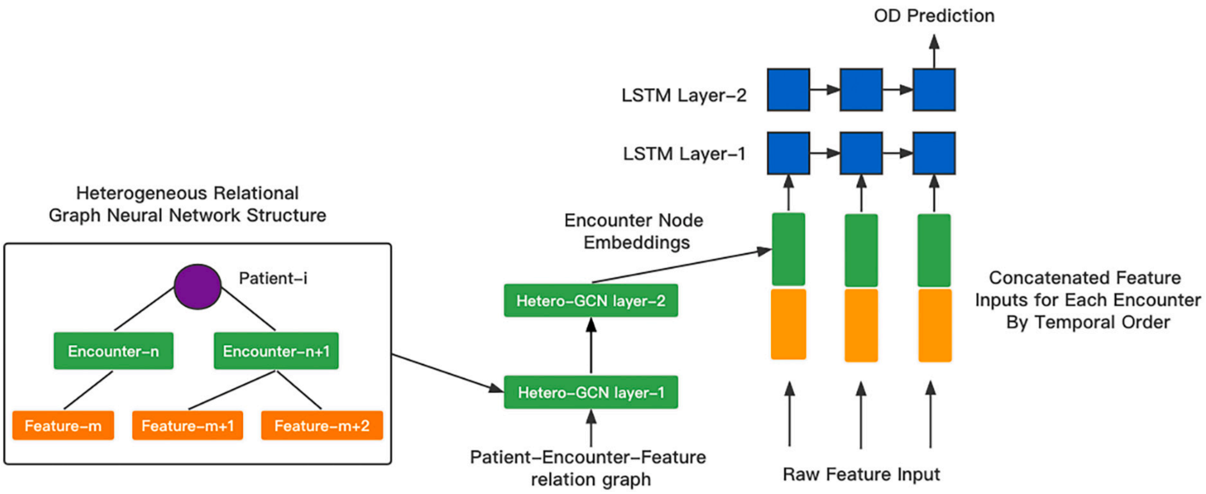
OD Prediction
LSTM Layer–2
LSTM Layer–1
Heterogeneous Relational Graph Neural Network Structure
Encounter Node Embeddings
Patient–i
Encounter–n
Encounter–n+1
Feature–m
Feature–m+1
Feature–m+2
Hetero–GCN layer–2
Hetero–GCN layer–1
Concatenated Feature Inputs for Each Encounter By Temporal Order
Patient–Encounter–Feature relation graph
Raw Feature Input

Fig. 6. Structure of integrated LSTM-GNN model.

graphic relation on the input features, every input feature is constructed as a numeric vector for a node or edge. Then a GNN model can be applied to those input features correlated by the graph structure to perform a prediction task. The node and edge vectors which represent the input feature are updated during the training phase. Finally, after training, the updated vectors are used as a graph embedding. Nodes or edges of similar structures in the graph will have similar embeddings [56], which means similar effects in terms of the prediction task. Thus the graph embeddings are naturally suited for grouping by a clustering algorithm. Fig. 7 shows the workflow we used to apply clustering algorithms on graph embeddings.

In our LIGHTED model, EHR features are processed as node vectors, and the machine learning task is opioid overdose prediction. After training, the heterogeneous relational graph node features are updated accordingly, and there is a vector of fixed length for each feature node that can be used as a graph embedding for the feature. The clustering of features' graph embeddings provides simplification of the feature set and reduces redundancy. Feature clustering may be more usable for clinical interpretations, since the grouped features in a feature's cluster may help to infer its meaning. Our resulting feature clusters were qualitatively evaluated by 2 board-certified clinicians to 1) assess clinically meaningful patterns and to potentially map the clusters to clinical topics, and 2) reassess the contents of the derived clusters for face-validity.

4. Experiment

4.1. Experiment setting

After selection, a dataset was built consisting of 5,231,614 patients with at least one opioid prescription. Among them, 44,774 patients had an opioid poisoning diagnosis (“positive”) and 5,186,840 patients did not have any recorded opioid poisoning event (“negative”). The total cohort of negative patients was portioned into 10 equal parts of 518,684 negative patients each. Each negative patient portion was combined with all the positive patients together for an evaluation, with random selection of 80 % as a training set and the remaining 20 % as the test set. We repeated the evaluation for all 10 portions, and an average value of each performance metric was reported for evaluation.

We calculated common machine learning performance metrics, including precision, recall, F1 score, and Area Under the Receiver Operating Characteristic curve (AUROC). Recall is a critical factor for prediction models as it identifies how many potential OD patients we can predict in advance. High recall can be achieved by tuning parameters to have a lower precision, since there is a tradeoff between recall and precision. Therefore, the F1 score, as a measurement considering both precision and recall, is regarded as the best aggregated assessment of the overall prediction performance. Along with the metrics, we also applied *t*-tests to compare the results between our proposed model and comparison methods to demonstrate that the novel model is statistically significantly better than existing models. We repeated the experiment 30 times, then applied a t-test to the sequence of F-1 scores and AUROC scores separately to compute the *p*-value.

Our implementation environment was the Python programming language (2.7). Traditional machine learning methods were implemented with the Python Scikit-Learn package [38]. Deep learning was implemented with Python TensorFlow [39], Python Keras [40] and PyTorch Geometric [41]. Other libraries used included Python NumPy [42], Python Pandas [43], and Python SHAP [44]. The training was performed on an NVIDIA Tesla V100 (16GB RAM).

4.2. Prediction result

We compared our models with other machine learning methods to evaluate the effectiveness of the proposed model. The comparison methods can be classified into four categories: 1) traditional methods (random forest, decision tree, logistic regression and dense neural network) are applied on raw input matrix without graph embeddings; 2) sequential methods (LSTM, bidirectional LSTM [45], attention model [46] and transformer [47] are applied on raw input without graph embeddings; and 3) graph models (GCN [12], GAT [13] and HeteroRGCN [16]) are applied on the graph embedding without raw input matrix; and 4) sequential graph combined models (LSTM-GNN, LSTM-GCN and LSTM-GAT) are applied on both graph embedding and raw input feature matrix in the same way as LIGHTED (LSTM-HeteroRGNN) but with a homogeneous graph. Table 2 shows the results for all compared methods.

The LIGHTED model achieved an F1 score of 0.8006. For precision, the best performance the LIGHTED model achieved was 0.8182, which indicates the accuracy of a positive OD prediction for a patient. The LIGHTED model also achieved the best recall, 0.7865, which measures the ability to find all the OD patients in the dataset. In terms of F1 score, the LIGHTED model had better performance than comparison methods. For the transformer model, the difference was not statistically significant (p-value$\geq$0.05). As for AUROC, the LIGHTED model performed significantly better than the transformer model, in addition to outperforming all other models.

Fig. 8(A) shows the ROC curves for 6 major methods: random forest, dense neural network, LSTM, Transformer, HeteroRGCN, and LIGHTED.

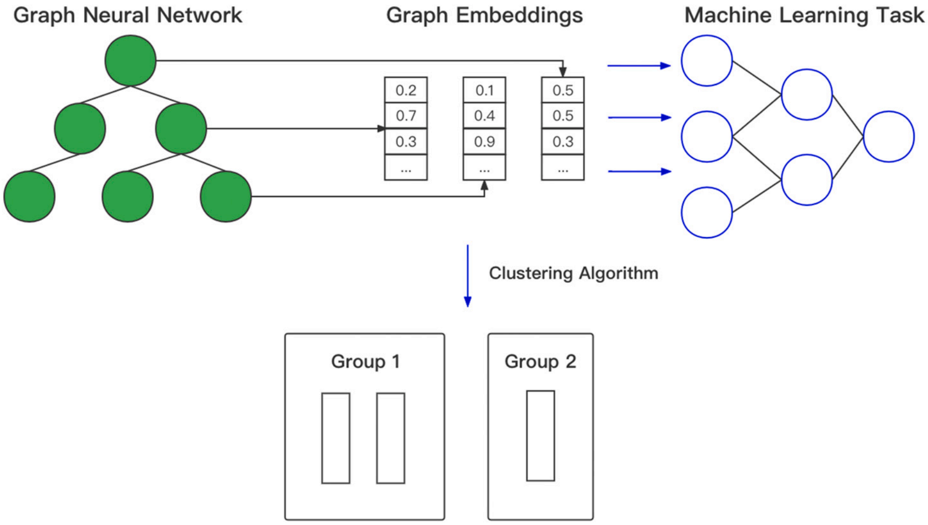

Fig. 7. The workflow of applying clustering algorithms on graph embeddings.

Table 2
Summary of prediction performance results.

Model	Precision	Recall	F-1	AUROC	p-value (F-1)	p-value (AUROC)
Traditional Methods (on raw feature input matrix)						
Random Forest	0.7695 ± 0.0056	0.7055 ± 0.0038	0.7361 ± 0.0057	0.8291 ± 0.0037	<0.01	<0.01
Decision Tree	0.7277 ± 0.0053	0.7047 ± 0.0029	0.7160 ± 0.0029	0.7892 ± 0.0048	<0.01	<0.01
Logistic Regression	0.7539 ± 0.0029	0.6050 ± 0.0026	0.6647 ± 0.0025	0.7147 ± 0.0035	<0.01	<0.01
DNN	**0.8006 ± 0.0052**	**0.7329 ± 0.0046**	**0.7683 ± 0.0027**	**0.8414 ± 0.0028**	<0.01	<0.01
Sequential Models (on raw feature input matrix)						
LSTM	0.7884 ± 0.0054	0.7616 ± 0.0027	0.7798 ± 0.0060	0.8618 ± 0.0051	<0.01	<0.01
Bi-LSTM	0.7879 ± 0.0013	0.7615 ± 0.0012	0.7796 ± 0.0012	0.8593 ± 0.0065	<0.01	<0.01
Attention	**0.8128 ± 0.0019**	0.7512 ± 0.0012	0.7815 ± 0.0022	0.8749 ± 0.0024	<0.01	<0.01
Transformer	0.8124 ± 0.0086	**0.7654 ± 0.0109**	**0.7911 ± 0.0019**	**0.8766 ± 0.0060**	0.117	<0.01
Graph Models (on graph embeddings)						
GCN	0.7867 ± 0.0089	0.7533 ± 0.0056	0.7696 ± 0.0020	0.8395 ± 0.0082	0.046	<0.01
GAT	0.7831 ± 0.0016	0.7433 ± 0.0086	0.7580 ± 0.0036	0.8292 ± 0.0095	<0.01	<0.01
HeteroRGCN	**0.8003 ± 0.0060**	**0.7679 ± 0.0021**	**0.7750 ± 0.0017**	**0.8429 ± 0.0046**	<0.01	<0.01
Sequential Graph Combined Models (on both raw input and graph embeddings)						
LSTM-GNN	0.7884 ± 0.0094	0.7728 ± 0.0096	0.7851 ± 0.0079	0.8589 ± 0.0133	<0.01	<0.01
LSTM-GCN	**0.7991 ± 0.0093**	**0.7767 ± 0.0088**	**0.7877 ± 0.0052**	0.8608 ± 0.0085	<0.01	<0.01
LSTM-GAT	0.7971 ± 0.0076	0.7692 ± 0.0112	0.7828 ± 0.0047	**0.8667 ± 0.0076**	<0.01	<0.01
Proposed Model						
LIGHTED (LSTM-HeteroRGNN)	**0.8182 ± 0.0072**	**0.7856 ± 0.0103**	**0.8006 ± 0.073**	**0.8969 ± 0.0115**	[*]	

The best performance of each metric in every category is marked in bold.
[*] *p*-values are generated by comparing each method with LIGHTED, so there are no p-values in the row of LIGHTED itself.

Fig. 8(B) shows how sensitivity and specificity change over different score thresholds for binary classification in LIGHTED. Fig. 8(C) shows the average time needed for one epoch when training each model. The LIGHTED model requires longer than common sequential and graph models, but shorter training time than the transformer model.

4.3. Risk scoring

To support clinical decision, model should be able to give an estimate of the level of risk beyond a binary prediction. The clinical practices to evaluate risk level of OD in the US are largely based on guidance from the CDC, where risk assessment is primarily driven by findings from the literature. While some tools have recently been developed for prediction of OD risk, they are mostly based on MME. For example, in the 2019 CMS opioid safety measures [57], patients of high OD risk are identified by high dose opioid use and concurrent use of benzodiazepines. Overall, those tools are limited to make assessments on MME and a few related medications. Given the paucity of tools for clinical practitioners in assessing risk of overdose, we suggest several methods for grouping patients into high/medium/low risk categories.

Based on the patients' probability scores from LIGHTED, we used four different methods to risk-stratify and group patients [32]. First, we split the test dataset into two equal parts, used the first part to decide the probability threshold to split the different risk level groups, then applied the thresholds to the second part to report the negative and positive rates for each risk level group to validate the performance. For the high risk group, we ranked patients by probability scores and then identified those in the percentile <1 % as high risk patients. We defined medium risk as those patients with probability scores in the 1 % to <10 % percentile, with the remaining patients designated as low risk. As a second method for scoring patient risk, we set the probability threshold that maximizes the F1 score as the threshold for the medium risk group. Our third method for risk scoring was to take the probability threshold when the summation of the sensitivity and specificity was maximized for the medium risk group [48]. For the fourth method, we took the probability when specificity equals 0.9 as the threshold for the medium risk group. For all the medium risk patients, we excluded the patients in the high risk group, while for the low risk group, we excluded the patients in high and medium risk groups. As a comparison, we applied the traditional clinical practice CMS opioid safety measure to split the risk group level. Table 3 shows the portion of true positive and true negative patients in each group after applying each grouping method.

In the high risk group, >96 % of patients are positive. For the medium risk group, around 60 % of patients are positive across the different methods. Compared to the clinical practice, CMS opioid safety measure, we have a low chance of misclassifying positive patients as negative. In the low risk group, only around 5 % of patients are positive. The ability to accurately identify risk groups can help clinicians and policymakers to better target interventions toward patients at high risk of OD.

4.4. Interpretation

To provide better interpretability of the LIGHTED model, we first provided and compared three importance feature ranking methods. We then proposed a novel interpretability approach by grouping features and patients with the graph neural network representations.

4.4.1. Feature importance ranking

We applied population level feature ranking methods to our proposed LIGHTED method, with the top 50 ranked features for each method listed in Table 4. We used permutation feature importance, a model-agnostic decision tree, and Shapley values to rank features. For the permutation method, we defined feature importance as the decreased value in AUROC metric when the value of that feature was blinded to the model [19]. Model-agnostic methods include the use of interpretable surrogate models such as a decision tree to simulate black-box deep learning models; the intrinsically interpretable model is trained using the black box model predictions instead of the outcomes from the original data set. The surrogate model output is then used to rank the features [17]. Shapley values involve a game theory-based

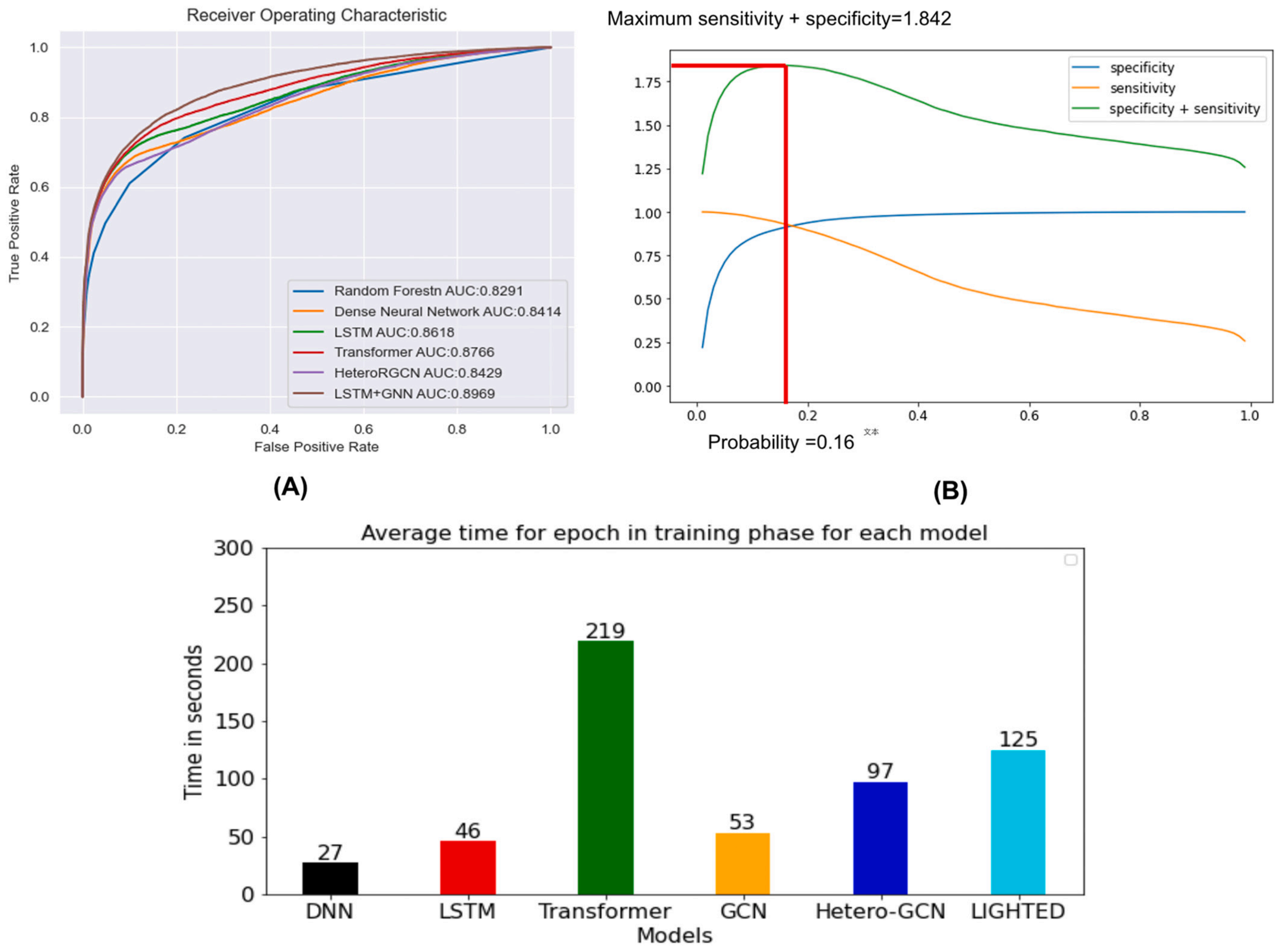

Fig. 8. (A) ROC curves for different models. (B) Specificity and sensitivity curves of LIGHTED over different probability thresholds. (C) Average time for epoch in training phase for each model.

Table 3
True positive and negative ratios by different high/medium/low grouping rules.

Risk Group	High Risk		Medium Risk (exclude high risk)		Low Risk (exclude high and medium risk)	
	Positive (%)	Negative (%)	Positive (%)	Negative (%)	Positive (%)	Negative (%)
Grouping Method	<1 percentile (1216)		>1 and < 10 percentile (4956)		other (50,174)	
	96.65	3.35	60.69	39.31	6.96	93.04
	<1 percentile (1216)		Maximum F1 score (4708)		other (50,422)	
	96.65	3.35	62.11	37.89	7.09	92.91
	<1 percentile (1216)		Maximum Sensitivity + Specificity (6528)		other (48,602)	
	96.65	3.35	52.48	47.52	6.37	93.63
	<1 percentile (1216)		Specificity = 0.9 (5268)		other (49,862)	
	96.65	3.35	58.81	41.19	6.83	93.16
CMS Opioid Safety Measure[a]	High Risk Group (10,049) Positive: 34.30 %, Negative: 65.7 %			Low Risk Group (46,297) Positive: 2.22 %, Negative: 97.78 %		

[a] The 2019 CMS opioid safety measures are meant to identify high-risk individuals or utilization behavior [57]. These measures include 3 metrics: (1) high-dose use, defined as higher than 120 morphine milligram equivalent (MME) for 90 or more continuous days, (2) 4 or more opioid prescribers and 4 or more pharmacies, and (3) concurrent opioid and benzodiazepine use for 30 or more days.

Table 4
Top 50 features for each interpretation method on LIGHTED model. Features with clinical relevance are labeled with a number, (1) features related to pain/opioids/drug misuse, (2) features related to a mental health disorder (3) features related to the respiratory system.

Rank	Shapley Value	Model-Agnostic Method (Decision Tree)	Permutation Importance
1	N02A: Opioids (1)	N02A: Opioids (1)	N02A: Opioids (1)
2	Other and unspecified disorders of back (1)	Blood Pressure Diastolic	Pain Scale Score (1)
3	Nondependent abuse of drugs (1)	MME (1)	Antipropulsives
4	MME (1)	Anesthetics, general	Anesthetics, general
5	Weight	Antipropulsives	Alcohol Use (1)
6	Carbon dioxide (3)	Respiratory Rate (3)	Other analgesics and antipyretics (1)
7	Essential hypertension	Other analgesics and antipyretics (1)	Blood Pressure Diastolic
8	BSA, Body Surface Area	Height	Blood Pressure Systolic (1)
9	Aspartate Aminotransferase / SGOT	Heart Rate	Hypnotics and sedatives (2)
10	General symptoms	BMI, Body Mass Index	Mean Corpuscular Hemoglobin
11	Respiratory Rate (3)	Weight	Red Blood Cell Distribution Width (RDW)
12	Diabetes mellitus	Pulse	MME (1)
13	Other analgesics and antipyretics (1)	Mean Arterial Pressure	Smoke, Exposure to Tobacco Smoke (1)
14	Other disorders of soft tissues	Temperature Oral	Uterotonics
15	Creatinine, Serum Quantitative	O2 Saturation (SO2) (3)	Blood Urea Nitrogen
16	Blood Urea Nitrogen	Uterotonics	Alkaline Phosphatase, Serum
17	Glucose, Serum/Plasma Quantitative	Red Blood Cell Distribution Width (RDW)	Heart Rate
18	Occupant of pick-up truck or van injured in noncollision transport accident (1)	Pulse Peripheral	Chloride, Serum
19	Red Blood Cell Distribution Width (RDW)	SPO2 (Saturation of peripheral oxygen) (3)	Height
20	Antiemetics and antinauseants	Heart Rate Monitored	Weight
21	Other symptoms involving abdomen and pelvis	Blood Urea Nitrogen	BMI, Body Mass Index
22	Anxiety, dissociative and somatoform disorders (2)	Temperature (Route Not Specified)	Monocyte Count
23	Hemoglobin	Weight, Pounds	Throat preparations
24	Other and unspecified disorders of joint (1)	Cough suppressants, excluding combinations with expectorants	Tobacco Use (Number of Years) (1)
25	Pain, not elsewhere classified (1)	White Blood Cell Count	Albumin, Serum
26	Symptoms involving respiratory system and other chest symptoms (3)	Pulse Oximetry (3)	Lymphocyte Absolute Count
27	Pain Scale Score (1)	Glasgow Coma Score (3)	Basophils Percent
28	Weight, Ideal	Weight, Clinical	Other and unspecified disorders of back (1)
29	Hematocrit	Hypnotics and sedatives (1)	Hemoglobin

Table 4 (*continued*)

Rank	Shapley Value	Model-Agnostic Method (Decision Tree)	Permutation Importance
30	White Blood Cell Count	Carbon dioxide (3)	Nondependent abuse of drugs (1)
31	Mean Platelet Volume	Mean Corpuscular Hemoglobin	Red Blood Cell Count
32	SPO2 (Saturation of peripheral oxygen) (3)	Weight, Daily Kilograms	Glomerular Filtration Rate Estimated
33	Anion Gap	Lymphocyte Percent	Posterior pituitary lobe hormones
34	Mean Corpuscular Hemoglobin	Posterior pituitary lobe hormones	BSA, Body Surface Area
35	Chloride, Serum	Weight, Ideal	Potassium, Serum
36	Glasgow Coma Score (3)	Glomerular Filtration Rate Estimated	Hematocrit
37	Platelet Count	Temperature Temporal Artery	Pain, not elsewhere classified (1)
38	O2 Saturation (SO2) (3)	Creatinine, Serum Quantitative	Calcium, Serum
39	Calcium, Serum	Mean Corpuscular Hemoglobin Concentration	Creatinine, Serum Quantitative
40	Episodic mood disorders (2)	Height, Inches	Anxiolytics (2)
41	Other postprocedural states	Chloride, Serum	Blood Gas CO2 Total, Arterial (3)
42	Drug dependence (1)	Alkaline Phosphatase, Serum	Lymphocyte Percent
43	Potassium, Serum	Glucose, Serum/Plasma Quantitative	Drug dependence (1)
44	Prothrombin Time	ETCO2 (End Tidal CO2) (3)	UA White Blood Cell
45	Symptoms involving digestive system	Albumin, Serum	Irrigating solutions
46	Albumin, Serum	Mean Platelet Volume	Essential (Primary) Hypertension
47	INR (International Normalized Ratio)	BSA, Body Surface Area	Neutrophil Percent
48	Sodium, Serum (3)	Occupant of pick-up truck or van injured in noncollision transport accident (1)	Erythrocytes Blood Automated Count
49	Chronic airway obstruction, not elsewhere classified (3)	Neutrophil Count	Ammonia
50	Blood Pressure Diastolic	Potassium, Serum	Pulse

approach to explain the output of any machine learning model [44]. A Shapley value measures the contribution of a given feature value to the difference between the actual prediction and the mean prediction. After calculating the Shapley value for every observed value of a given feature, the results are aggregated by taking the mean (irrespective of direction); this aggregate mean is compared across features to form an importance ranking.

After ranking the features according to each of the three methods (model-agnostic decision tree, Shapley value, and permutation importance), a practicing clinician investigator used knowledge of clinically-recognizable concepts and a priori features with a known OD and/or opioid use disorder association to qualitatively evaluate the top 50 features from each interpretability method for their potential clinical relationships to opioid overdose. The characterization of features related to OD are shown in Table 4. Of the 3 methods, the top 50 features identified using Shapley values (shown in Fig. 9) included the most clinically meaningful features (20), with 11 features related to pain/opioid/non-medical drug use (marked with (1) in Table 4, yellow bars in Fig. 8), 2 features corresponding to mental disorders (marked with (2) in Table 4, blue bars in Fig. 8) [49], and 7 corresponding to features affecting the respiratory system (marked with (3) in Table 4, green bars in Fig. 8) [50].

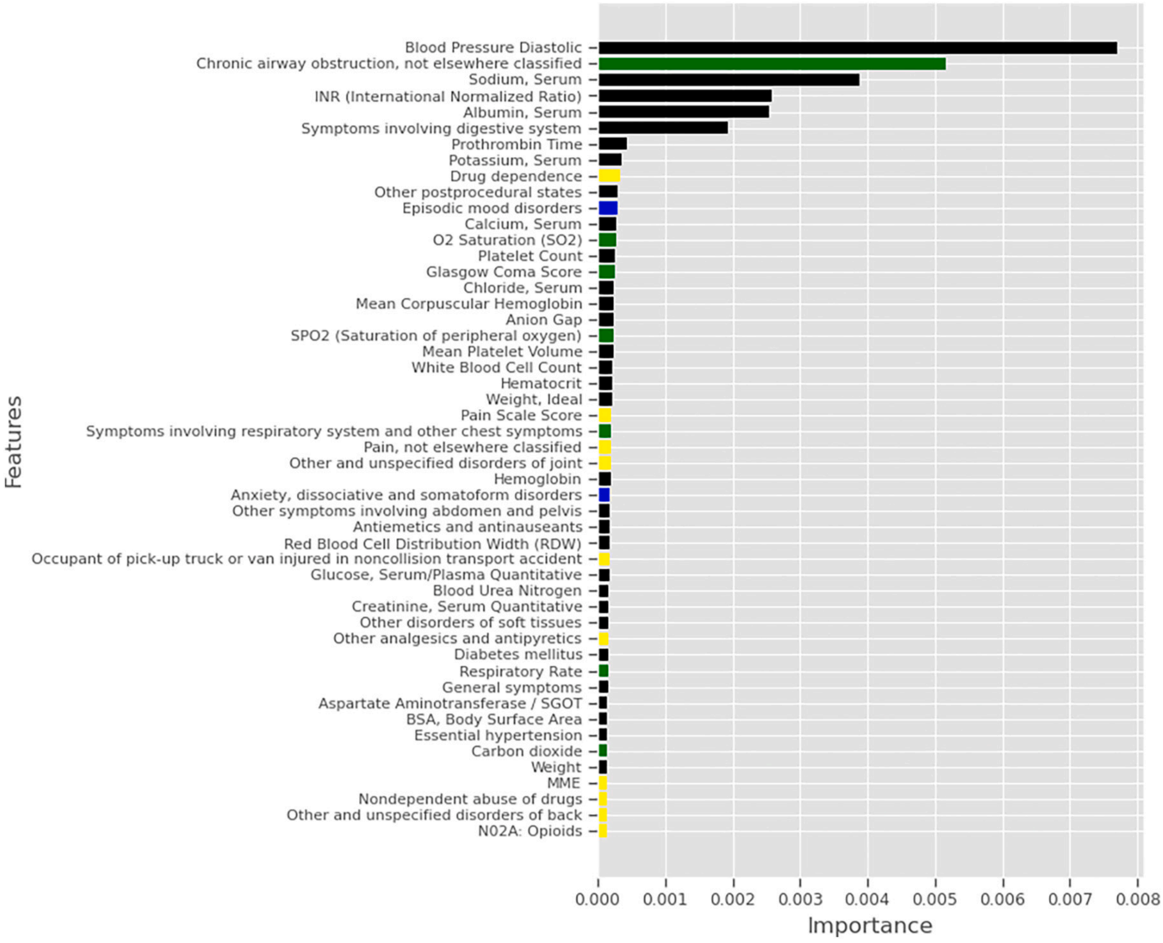

Fig. 9. Top 50 important features identified by Shapley value. Features in yellow are related to pain/opioid/drug misuse, features in blue are related to mental disorders and features in green are related to respiratory system. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.4.2. *Feature clustering*

To potentially simplify the feature space and reduce concept redundancy, we ran an agglomerative hierarchical clustering algorithm on the graph embeddings of all features. Clusters were evaluated by a clinical investigator to assess for clinically meaningful patterns where at least 25% of the features had a definable clinical relationship to each other and thus gave the cluster potential explainability. Table 5 shows partial results of feature clustering. Some clusters (A, B) were highly interpretable in the context of feature cluster relationship to OD, such as clusters with multiple diagnoses related to accidental injury or substance/medication use or toxicity [51,52]. Other clusters (C), which we defined as intermediate in explainability, contained interpretable features related to clinical reno-pulmonary systems. They may be related to OD through mechanisms such as respiratory depression or impaired renal function [53,54]. Some clusters (D) did not have any clearly identifiable patterns with clinical meaning.

5. Discussion

With the wide availability of electronic health records, predictive modeling can provide a powerful approach to estimate risks of opioid overdose for patients and support early interventions for prevention and risk mitigation. However, the complex dynamics and interactions of EHR data make it difficult for traditional machine learning and statistics-based predictive models to process. Advanced deep learning models can address these challenges of complex EHR data. Specifically, the sequential LSTM can model the dynamics of EHR data, and graph neural networks can model the interactions between EHR data. Therefore, we developed a new integrated model, LIGHTED, combining both LSTM and GNN frameworks to better address the opioid overdose risk prediction problem.

Our proposed LIGHTED integrated model achieved a promising result, with a higher F1 score than traditional learning models, sequential models, or graph models alone. To improve model interpretability, we identified top features related to opioid overdose, which could be used as explanatory mechanisms to support clinicians in the risk assessment process, and potentially, in a toolkit for clinical decision support. We also explored feature embeddings provided by the graph clustering algorithm to group features into feature subsets with clinical meaning, as a potential simplified representation of high dimensional EHR features.

5.1. *Benefits of the model*

Compared with other traditional and deep learning models, our proposed model has several advantages. First, we can model the dependence relationship between observations at different time steps. Second, the interactions between features and patients can be learned from the heterogeneous graph based model. Third, graph embeddings can be used to cluster features and glean higher-level interpretations of the features when clinically meaningful patterns emerge.

Table 5
Samples of clustering result on graph embeddings.[a]

Cluster A. Accidental Injury (High*)	Cluster B. Substance/Medication Use or Toxicity (High*)
Other slipping, tripping and stumbling and falls	Acetaminophen, serum quantitative
Occupant of pick-up truck or van injured in collision with fixed or stationary object	Alcohol and/or drug, substance use
Occupant of pick-up truck or van injured in noncollision transport accident	Alcohol last use
Occupant of pick-up truck or van injured in collision with heavy transport vehicle or bus	Tobacco frequency other
Other complications of procedures not elsewhere classified	Tobacco last use
Motor vehicle traffic accidents	Smoking, attempt to quit in Past
Accidents caused by submersion, suffocation, and foreign bodies	Smoking, willing to quit
Other disorders of bone and cartilage	Smoking, readiness to quit
Contusion of trunk	Smoke, lives with someone who smokes
Glasgow Coma Score	Smoking packs/day
BSA, body surface area estimated	Smoke, exposure to tobacco smoke
Abdominal and pelvic pain	Smoking history
Injury of unspecified body region	Tobacco type
Injury other and unspecified	Opioid related disorders
Drugs, medicinal and biological substances causing adverse effects in therapeutic use	Pain scale score
MME	Muscle relaxants, peripherally acting agents
Other personal history presenting hazards to health	Drugs for constipation
Pain, not elsewhere classified	Barbiturate, urine
Pain associated with micturition	Poisoning by psychotropic agents
Episodic mood disorders	Antipsychotics
Personality disorders	Personal history of mental disorder
Phencyclidine Urine	QT corrected (QTc)

Cluster C. Reno-Pulmonary Features (Intermediate*)	Cluster D. No Identifiable Pattern (Low)
Essential (primary) hypertension	Vitamin B1, plain and in combination with vitamin B6 and B12
Blood pressure diastolic sitting	Red blood cell distribution width (RDW)
Blood pressure systolic sitting	Alcohol use
Glomerular filtration rate/1.73 sq. M predicted among blacks creatinine based formula (MDRD)	Blood pressure central venous
UA bacteria	Neutrophil segmented percent
Protein, urine	Intestinal anti-inflammatory agents
Protein total, urine random	Special screening examination for bacterial and spirochetal diseases
Potassium, whole blood	Antivaricose therapy
Calcium, serum	Other systemic drugs for obstructive airway diseases
Bicarbonate HCO3	Disorders of external ear
HCO3	Encounter for immunization
Other diseases of lung	LDL/HDL ratio
Other diseases of respiratory system	Gastritis and duodenitis
Drugs for treatment of tuberculosis	Unspecified viral hepatitis
Symptoms involving respiratory system and other chest symptoms	Carboxyhemoglobin
Pain in throat and chest	Osteoarthrosis and allied disorders
PIP (Peak inspiratory pressure)	Symptoms involving urinary system

[a] * High/Medium/Low is the strength of connection between the features and their cluster titles, according to two board-certified clinicians.

5.2. *Clinical significance*

It is critical for OD prevention to identify high risk patients and learn what features are related to the development of OD. Our model provides the prediction for each patient and meaningful grouping of patients by the risk level. Moreover, it identifies EHR features most relevant to OD risk. To evaluate which features were most important for prediction across different models, we identified 13 features that were included in the top 50 features in all three models. Some of these features were directly related to pain or opioid related treatment, such as MME and medication category N02A: Opioids and Other analgesics and antipyretics. Three additional features may have an indirect influence on risk of OD: high Creatinine, Serum Quantitative; low Albumin; and low Weight. These features clinically correspond to patients with either impaired renal function, poor nutritional status or poor synthetic liver function, which could increase the risk of OD because of impaired ability to metabolize opioids. Other features did not have a recognizable mechanistic correlation with risk of OD. For the clustering on graph embeddings of features, we can see that some groups show high interpretability for opioid overdose patient identification. With those groups of features with clear clinical meaning, simplified explanatory labels can be given to represent the corresponding clinical concept, increasing the explainability for clinicians and patients.

5.3. *Limitations and future work*

Study limitations include that the study population was extracted from patients that received opioid prescriptions, so the results may not generalize well to patients who used non-prescribed opioids. Second, since the ground truth of the study was based on ICD codes in electronic health records, there are potentially patients who had OD at home or who were coded incorrectly that may have been missed. To address those problems, we plan apply natural language processing methods to incorporate unstructured notes data in our model, since structured EHR data may have missing information. Third, our model is based on data from the Health Facts database, which may not be applied to other EHRs directly. To address this issue, in future studies, we will exploit transfer learning methods to integrate knowledge learned from the Health Facts data with EHR data from our local health system.

6. Conclusion

Opioid overdose has become a leading cause of accidental death in the United States, which requires pharmacologic interventions and health efforts to address the epidemic. Predicting patients with high risk for OD can guide the early interventions in the developmental trajectory. Our integrated LSTM-HeteroRGNN model LIGHTED showed promising results on prediction of high risk patients, and provided features and feature clusters that are clinically meaningful to OD risk.

Declaration of competing interest

The authors do not have any conflicts of interest to disclose.

Acknowledgements

This work was funded partially by the Stony Brook University OVPR Seed Grant 1158484-63845-6 and National Institutes of Health 5R01CA253368.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2022.102439.

References

[1] Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and Health. Rockville: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2020.

[2] Center for Disease Control and Prevention. Assessing and Addressing Opioid Use Disorder (OUD). Available from:. 2020. https://www.cdc.gov/drugoverdose/training/oud/index.html.

[3] CDC/NCHS. National vital statistics system, mortality. Available from: https://wonder.cdc.gov; 2018.

[4] Che C, Xiao C, Liang J, Jin B, Zho J, Wang F. An rnn architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In: Proceedings of the 2017 SIAM international conference on data mining. Society for Industrial and Applied Mathematics; 2017. p. 198–206. https://doi.org/10.1137/1.9781611974973.23.

[5] Pillai NS, Bee KK, Kiruthika J. Prediction of heart disease using rnn algorithm. International research journal ofEngineering and Technology 2019:5.

[6] Cui R, Liu M, Li G. Longitudinal analysis for Alzheimer's disease diagnosis using RNN. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018); 2018. p. 1398–401. https://doi.org/10.1109/ISBI.2018.8363833.

[7] Dong X, Deng J, Hou W, Rashidian S, Rosenthal RN, Saltz M, Saltz JH, Wang F. Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning. J Biomed Inform 2021 Apr;1(116):103725. https://doi.org/10.1016/j.jbi.2021.103725.

[8] Dong X, Rashidian S, Wang Y, Hajagos J, Zhao X, Rosenthal RN, Kong J, Saltz M, Saltz J, Wang F. Machine learning based opioid overdose prediction using electronic health records. In: AMIA Annu Symp Proc; 2020. p. 389–98. PMID: 32308832; PMCID: PMC7153049.

[9] Dong Xinyu, Deng Jianyuan, Rashidian Sina, Abell-Hart Kayley, Hou Wei, Rosenthal Richard N, Saltz Mary, Saltz Joel H, Wang Fusheng. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. J Am Med Inform Assoc 2021;28(8):1683–93. https://doi.org/10.1093/jamia/ocab043.

[10] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw 2009;20(1):61–80. https://doi.org/10.1109/TNN.2008.2005605.

[11] Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 2021;32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386.

[12] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Information Process Syst 2016;29:3844–52.

[13] Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint; 2017. arXiv:1710.10903.

[14] Zhang Chuxu, Song Dongjin, Huang Chao, Swami Ananthram, Chawla Nitesh V. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (KDD '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 793–803. https://doi.org/10.1145/3292500.3330961.

[15] Wanyan Tingyi, Honarvar Hossein, Azad Ariful, Ding Ying, Glicksberg Benjamin S. Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. DataIntelligence 2021;3(3):329–39. https://doi.org/10.1162/dint_a_00097.

[16] Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: The semantic web. ESWC 2018. lecture notes in computer science. 10843. Cham: Springer; 2018. https://doi.org/10.1007/978-3-319-93417-4_38.

[17] Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint; 2016. arXiv:1606.05386.

[18] Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16). New York, NY, USA: Association for Computing Machinery; 2016. p. 1135–44. https://doi.org/10.1145/2939672.2939778.

[19] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.

[20] University of Texas Health Science Center at Houston. SBMI Data Service. Accessed September 22, 2020. https://sbmi.uth.edu/sbmi-data-service/data-set/cerner/.

[21] Wishart David S, Feunang Yannick D, Guo An C, Lo Elvis J, Marcu Ana, Grant Jason R, Sajed Tanvir, Johnson Daniel, Li Carin, Sayeeda Zinat, Assempour Nazanin, Iynkkaran Ithayavani, Liu Yifeng, Maciejewski Adam, Gale Nicola, Wilson Alex, Chin Lucy, Cummings Ryan, Le Diana, Pon Allison, Knox Craig, Wilson Michael. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research 2018;46(D1):D1074–82. https://doi.org/10.1093/nar/gkx1037.

[22] Moore B, Barrett M. Case study: exploring how opioid-related diagnosis codes translate from ICD-9-CM to ICD-10-CM2017; 2018.

[23] American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013.

[24] Deng Jianyuan, Yang Zhibo, Ojima Iwao, Samaras Dimitris, Wang Fusheng. Artificial intelligence in drug discovery: applications and techniques. Briefings in Bioinformatics 2022;23(1):bbab430. https://doi.org/10.1093/bib/bbab430.

[25] Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain—United States, 2016. JAMA 2016;315(15):1624–45. https://doi.org/10.1001/jama.2016.1464.

[26] Centers for Disease Control and Prevention. Screening list of ICD-9-CM codes for casefinding. https://www.cdc.gov/cancer/apps/ccr/icd9cm_codes.pdf (accessed March 21, 2021).

[27] Centers for Disease Control and Prevention. ICD-10-CM table of NEOPLASMS. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2019/icd10cm_neoplasm_2019.pdf (accessed March 21, 2021).

[28] Fareed A, Stout S, Casarella J, Vayalapalli S, Cox J, Drexler K. Illicit opioid intoxication: diagnosis and treatment. Subst Abuse Res Treat 2011:5. https://doi.org/10.4137/SART.S7090.

[29] Haghpanah T, Afarinesh M, Divsalar K. A review on hematological factors in opioid-dependent people (opium and heroin) after the withdrawal period. Winter–Spring Addict Health 2010;2(1-2):9–16. PMID: 24494095; PMCID: PMC3905505.

[30] Guzel Derya, Yazici Ahmet Bulent, Yazici Esra, Erol Atila. Evaluation of immunomodulatory and hematologic cell outcome in heroin/opioid addicts. Journal of Addiction 2018;2018:2036145. https://doi.org/10.1155/2018/2036145. 8 pages.

[31] Becker Daniel E, Phero James C. Drug therapy in dental practice: nonopioid and opioid analgesics. Anesth Prog 2005;52(4):140–9. https://doi.org/10.2344/0003-3006(2005)52[140,DTD]2.0.CO;2.

[32] Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. J JAMA Network Open 2019;2(3):e190968-e. PMCID: PMC6583312.

[33] White AG, Birnbaum HG, Schiller M, Tang J, Katz NP. Analytic models to identify patients at risk for prescription opioid abuse. Am J Manag Care 2009;15(12):897–906. PMID: 20001171.

[34] Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. Pain Med 2012;13(9):1162–73. https://doi.org/10.1111/j.1526-4637.2012.01450.x.

[35] Oliva EM, Bowe T, Tavakoli S, et al. Development and applications of the veterans health Administration's stratification tool for opioid risk mitigation (STORM) to improve opioid safety and prevent overdose and suicide. Psychol Serv. 2017;14(1):34–49. https://doi.org/10.1037/ser0000099.

[36] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

[37] Breiman Leo. Statistical modeling: the two cultures (with comments and a rejoinder by the author). StatSci 2001;16(3):199–231. https://doi.org/10.1214/ss/1009213726.

[38] Pedregosa F, et al. In: Scikit-learn: machine learning in Python. 12; 2011. p. 2825–30.

[39] Abadi M, et al. Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on operating systems design and implementation ({OSDI} 16); 2016.

[40] Chollet F, et al. Keras. Available from: GitHub; 2015. https://github.com/fchollet/keras.

[41] Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. arXiv preprint; 2019. arXiv:1903.02428.

[42] Bressert E. SciPy and NumPy: an overview for developers. O'Reilly Media, Inc.; 2012.

[43] McKinney W. "pandas: a foundational Python library for data analysis and statistics". In: Python for high performance and scientific computing. 14; 2011. p. 1–9. dlr.de.

[44] Lundberg Scott M, Lee Su-In. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems (NIPS'17). Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–77.

[45] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81. https://doi.org/10.1109/78.650093.

[46] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: 31st conference on neural information processing systems; 2017. p. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[47] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint; 2018. arXiv:1810.04805.

[48] Fluss R, Faraggi D, Reiser B. Estimation of the youden index and its associated cutoff point. Biom J 2005;47:458–72. https://doi.org/10.1002/bimj.200410135.

[49] Turner BJ, Liang Y. Drug overdose in a retrospective cohort with non-cancer pain treated with opioids, antidepressants, and/or sedative-hypnotics: interactions with mental health disorders. J Gen Intern Med 2015;30:1081–96. https://doi.org/10.1007/s11606-015-3199-4.

[50] Nadpara Pramit A, Joyce Andrew R, Murrelle ELenn, Carroll Nathan W, Carroll Norman V, Barnard Marie, Zedler Barbara K. Risk factors for serious prescription opioid-induced respiratory depression or overdose: comparison of commercially insured and veterans health affairs populations. Pain Medicine 2018; 19(1):79–96. https://doi.org/10.1093/pm/pnx038.

[51] Asfaw A, Boden LI. Impact of workplace injury on opioid dependence, abuse, illicit use and overdose: a 36-month retrospective study of insurance claims. Occup Environ Med 2020;77(9):648–53.

[52] Barefoot Elizabeth H, Cyr Julianne M, Brice Jane H, Bachman Michael W, Williams Jefferson G, Cabanas Jose G, Herbert Kyle M. Opportunities for emergency medical services intervention to prevent opioid overdose mortality. Prehosp Emerg Care 2021;25(2):182–90. https://doi.org/10.1080/10903127.2020.1740363.

[53] Vu Q, Beselman A, Monolakis J, Wang A, Rastegar D. Risk factors for opioid overdose among hospitalized patients. J Clin Pharm Ther 2018;43:784–9. https://doi.org/10.1111/jcpt.12701.

[54] Fox LM, Hoffman RS, Vlahov D, Manini AF. Risk factors for severe respiratory depression from prescription opioid overdose. Addiction 2018;113:59–66. https://doi.org/10.1111/add.13925.

[55] Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural NetwLearn Syst 2022;33(2):494–514. https://doi.org/10.1109/TNNLS.2021.3070843.

[56] Grover Aditya, Leskovec Jure. Node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16). New York, NY, USA: Association for Computing Machinery; 2016. p. 855–64. https://doi.org/10.1145/2939672.2939754.

[57] Centers for Medicare & Medicaid Services (CMS). Announcement of calendar year (CY) 2019 Medicare Advantage capitation rates and Medicare Advantage and Part D payment policies and final call letter. Accessed November 6, 2018. https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/Announcement2019.pdf.