Patient ADE Risk Prediction through Hierarchical Time-Aware Neural Network Using Claim Codes

Jinhe Shi*, Xiangyu Gao*, Chenyu Ha†, Yage Wang†, Guodong Gao‡ and Yi Chen*

*New Jersey Institute of Technology, Newark, NJ, USA
Email: {js675, xg77, yi.chen}@njit.edu
†Inovalon, Bowie, MD, USA
Email: {cha, ywang2}@inovalon.com
‡University of Maryland, College Park, MD, USA
Email: ggao@rhsmith.umd.edu

Abstract—Adverse drug events (ADEs) are a serious health problem that can be life-threatening. While a lot of studies have been performed on detect correlation between a drug and an AE, limited studies have been conducted on personalized ADE risk prediction. Among treatment alternatives, avoiding the drug that has high likelihood of causing severe AE can help physicians to provide safer treatment to patients. Existing work on personalized ADE risk prediction uses the information obtained in the current medical visit. However, on the other hand, medical history reveals each patients unique characteristics and comprehensive medical information. The goal of this study is to assess personalized ADE risks that a target drug may induce on a target patient, based on patient medical history recorded in claims codes, which provide information about diagnosis, drugs taken, related medical supplies besides billing information. We developed a HTNNR model (Hierarchical Time-aware Neural Network for ADE Risk) that capture characteristics of claim codes and their relationship. The empirical evaluation show that the proposed HTNNR model substantially outperforms the comparison methods, especially for rare drugs.

Index Terms—Adverse Drug Event, Neural Network, Claim Code

I. INTRODUCTION

Adverse Drug Events (ADEs), defined as "an appreciably harmful or unpleasant event resulting from the use or misuse of a drug" [1], are a serious health problem that can be life-threatening. According to FDA, the number of ADEs reported to FAERS (FDA Adverse Event Reporting System) resulting in death and serious outcomes increase consistently [2], [3]. Statistics [4] show each year ADEs account for over 3.5 million physician office visits, an estimated 1 million emergency department visits, and approximately 125,000 hospital admissions. For inpatient setting, ADEs account for an estimated 1 in 3 of all hospital *adverse events (AE)* and affect about 2 million hospital stays each year.

While pre-marketing review is conducted before any drugs are approved for marketing, it is insufficient for identifying all the potential ADEs due to the limited sample size and duration of clinical trials. Post-marketing surveillance is critical for identifying ADRs. Although patient can report ADE through voluntary and spontaneous report systems, such as FDA FAERS, the median under-reporting rate across 37 studies

using a wide variety of post-marketing surveillance methods from 12 countries is 94% according to an earlier study [5].

There are increasing interests of using large-scale longitudinal clinical data, EHRs, associated clinical notes, as well as claims data, for studying ADEs. Such data contain rich and accurate information about patients health status, their treatment plan and clinical outcomes. Since such data is generated as part of medical practices, without relying on patient self-reporting, it is available in large-scale with high quality.

The studies can be categorized into two types: ADE detection and personalized ADE risk prediction. The goal of ADE detection is to identify the correlation or causal relationship between a target drug and an observed AE. Some use statistic methods such as the disproportionality analysis [6]–[8], others use machine learning methods such as support vector machines, random forests and neural networks [9], [10]. Besides detecting drug-AE correlation on the whole patient population, there are also studies on ADE risk stratification which assesses the correlation on patient populations defined by their demographics [11].

In contrast of ADE detection for population, the studies on personalized ADE risk prediction assess the likelihood of individuals to experience an AE based on individual characteristics and clinical history. Indeed different patient may have different AE outcomes even taking the same drug. Among alternative drugs for treatment, avoiding the one that has high likelihood of causing severe AE can help physicians to provide safer treatment to patients, as a form of personalized treatment. There are only a few works addressing the problem [12], [13]. They take as input patient demographic information and clinical information of the current hospital visit.

What is lacking in the literature is to consider patient medical history in addition to the current visit information to make personalized prediction for ADE risks. Medical history better reveals each patient's unique characteristics, as well as the drugs and treatments taken in the past, which may interact with the current treatment to induce AE [14], [15].

However, patient medical history data is often not readily available and is difficult to process. First, patient may be seen at multiple healthcare centers that do not share patient data

in their EHR system. Second, patient self-reporting medical history may not be accurate or comprehensive. Furthermore, processing large-scale longitudinal medical history data, which contains diverse type of clinic information, poses technical complexity

The goal of this study is to assess personalized ADE risks that a target drug may induce on a target patient, based on patient medical history recorded in claims, which we acquired access via collaboration with Inovalon, a healthcare analytics company. Our findings can be used by Medicare/Medicaid and health insurance company to provide assistance to healthcare professionals to identify safe treatment plan.

Claims data provide valuable information about patients. It contains the information about diagnoses, drugs taken, related medical supplies, treatment procedures, besides billing information, for each patient encounter. While a patient may receive healthcare from multiple providers, and have their medical information scattered in multiple EHR systems, claims data effectively records a patients interactions across different healthcare systems and thus provides longitudinal and accurate data in the continuum of a patient's health care history [16].

However, there are several technical challenges that must be addressed. The first challenge is how to capture the "meanings" of claim codes. There are over 64K unique claims code in the data, belong to nine different types. We make an analogy between claim codes and words, and between claim history and documents. Then we propose to use word embedding methods in Natural Language Processing (NLP) to generate embedding for claim codes, so that claim codes that are used in similar ways are represented with similar vectors, naturally capturing their meanings.

The second challenge is how to model patient medical claim history. A patient's claim history consists of encounters and each encounter consists of claim codes. The relationship of claim codes within an encounter is different from that of claim codes in different encounters. This present a unique challenge, as exiting work does not consider patient's medical history but only the current medical visit. To model patient's claim history, we propose a HTNNR model stands for Hierarchical Time-aware Neural Network with drug-code Representation. The first layer neural network encodes claim codes within an encounter into vectors, and the second layer neural network represents the claim history with a sequence of encounters into vectors. Then we propose to use a bi-directional neural network model to capture the un-ordered relationship among claim codes within an encounter. We further propose to use time-aware deep learning model to capture not only the sequential but also the temporal relationship among encounters.

The contributions of our work include the following. First, to the best of our knowledge, this the first study that uses patient claim history to make personalized prediction on drug-induced ADE risks. Second, we have made several technical contributions. We proposed claims code embedding, a hierarchical neural network model to capture patient claim history, and drug-claim code representations. We also used different neural network models for encounter representation and for claim history representations. Finally, extensive evaluation on about 500k patients demonstrates effective prediction performance and high efficiency of our proposed approach.

The rest of the paper is organized as follows. Section II discusses the related work. Section III presents the problem statement and data overview. Section IV and Section V presents the two methods for patient ADE risk prediction. Experimental results are presented in Section VI. Section VII concludes the paper.

II. RELATED WORK

Studies on ADEs can be categorized into ADE detection on population, personalized AE risk prediction, and prediction of ADE outcome intensity (e.g. hospitalization and mortality).

The goal of ADE detection is to identify the correlation or causal relationship between a target drug and an observed AE, using statistical methods or machine learning methods. Some studies applied association rule mining methods for ADE detection [17], [18]. Disproportionality analysis are widely used for ADE detection from various data sources, such as EHR data [19], [20], clinic notes [8], and clinical trials [21].

Disproportionality analysis is based on the contrast between observed and expected numbers of co-occurrences, for any given combination of drug and AE, to detect possible causal relations between drugs and AEs. It, however, does not consider context features, which are rich in unstructured clinical notes. Various Natural Language Processing (NLP) and machine learning techniques have been applied on clinic notes to detect drug-AE association, using expert-labeled ground truth. [9], [10] extract multiple features like drug and AE frequency and co-mention frequency from clinical notes and use machine learning methods like support vector machine and random forest to detect drug-AE correlation. [22], [23] start with a named entity recognition module based on Conditional Random Fields to extract medical entities relevant to ADEs from clinical notes, and then use random forest and neural networks, respectively, as the relation classification model. Little has been studied on using claims data for ADE detection. [24] use ICD codes and GPI drug code in claims data (see Table I for description of the code) as input and design a graph neural network model to construct a drug-disease graph for ADE detection. They first embedded disease codes and drug codes into a graph, respectively, then the merged drug and disease graph is fed into a graph neural network for ADE detection. They used the SIDER database as the ground truth for ADEs. Besides detecting drug-AE correlation on the whole patient population, there are also studies on ADE risk stratification which assesses the drug-AE correlation on patient populations defined by their demographics [11].

There are only a few studies in the category of personalized AE risk prediction. Since AE risks of different patients are different, even for the same drug, these studies make risk predictions based on the individual patient's characteristics from clinical data. [13] develops a logistic regression model to predict the risks of AEs of in-patients based on the patient features and the medical conditions during this hospital stay.

TABLE I
DESCRIPTION OF DIFFERENT CLAIM CODES

Code Type	Description
ICD	International Statistical Classification of Diseases (ICD) codes capturing diseases, symptoms, abnormal findings, complaints, etc. It includes diagnosis codes (ICD10DX and ICD9DX) and procedure codes (ICD9PX and ICD10PX).
CPT	report medical, surgical, and diagnostic procedures and services to entities such as physicians, health insurance companies and accreditation organizations
POS	Place of Service (POS) Codes are two-digit codes placed on health care professional claims to indicate the setting in which a service was provided.
GPI	The Generic Product Identifier (GPI) is a 14-character hierarchical classification system that identifies drugs from their primary therapeutic use down to the unique interchangeable product regardless of manufacturer or package size.
TOB	Type of bill codes (TOB) identifies the type of bill being submitted to a payer. TOB codes are four-digit alphanumeric codes that specify different pieces of information on claim form
REVENUE	Revenue Codes are descriptions and dollar amounts charged for hospital services provided to a patient.
HCPCS	The Healthcare Common Procedure Coding System (HCPCS) is a collection of codes that represent procedures, supplies, products and services which may be provided to Medicare beneficiaries and to individuals enrolled in private health insurance programs.
DISCHARGE	Identify where the patient is at the conclusion of a health care facility encounter (a visit or an inpatient stay)
LOINC	Logical Observation Identifiers Names and Codes (LOINC) is a database and universal standard for identifying medical laboratory observations

They used multiple patient characteristics like gender and age as features, also extracted some features from current medical conditions like the number of medications and the list of drugs taken. [12] takes clinical features as input, such as ADE indication codes, primary diagnosis code and length of the hospital stay to predict in-patient ADE risks. They used multiple machine learning models like random forest and support vector machines. Both make ADE risk prediction based on the information of the current hospital stay. Being most related to this category of studies, our work takes as input a patient's longitudinal medical history, not just the current medical encounter. Also, we consider AE risks induced by target drugs (perhaps due to interaction with other drugs or medical conditions), whereas existing studies consider AE in general. The dataset used in our studies is claims data.

Unlike studies on personalized AE risk prediction, which predict the likelihood of a specific AE to occur, there are also studies on predicting the likelihood of hospitalization and mortality of a patient, due to outcomes of unspecified AEs. Both of them are using the patient medical data from FAERS. [25] proposed a hybrid model to predict the outcomes of ADEs, based on patients demographic data, such as age and gender, and drug-taken information, such as the route of the drug intake and whether the adverse reaction subsided when drug in-take was terminated. [26] developed a system that takes patient demographics, drugs, relevant diseases in pathology as input, and outputs ADE risk outcome assessment.

III. PROBLEM STATEMENT

In this section we present the data description and the problem definition.

A. Data Description

The input data is medical claim history for a set of patients. Each claim history is composed of a sequence of encounters, and each encounter has a sequence of claim codes, as illustrated in Figure 3. At an encounter, a medical treatment and/or evaluation and management services are provided. There are nine different types of claim codes, which provide information of medical diagnoses (ICD), procedures and services (CPT, LOINC), setting where services are provided (POS), drug information (GPI), billing (TOB, REVENUE, DISCHARGE), and codes for Medicare and private health insurance program users (HCPCS). Table I shows a description about these code types.

The data used in empirical evaluation was provided by Inolvaon, a technology company providing cloud-based platforms empowering data-driven healthcare. It contains the claims data of 500k patients for a duration of 2015-2019. There are 64,070 unique claim codes. Figure 2 shows the distribution of the number of encounters a patient has. We can see that most patient has less than 500 encounters, and the average number of encounters per patient is 158.7. The average number of claim codes per patient is 1052 and the

average number of claim code per encounter is 6.6. Figure 1 shows the number of claim code occurrences of each category.
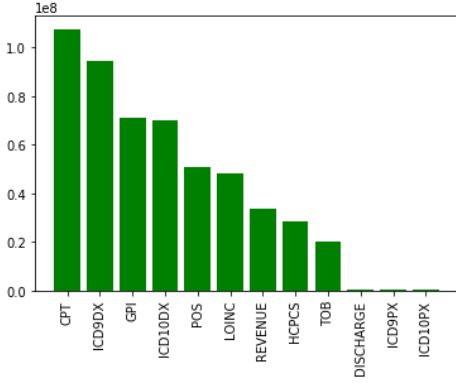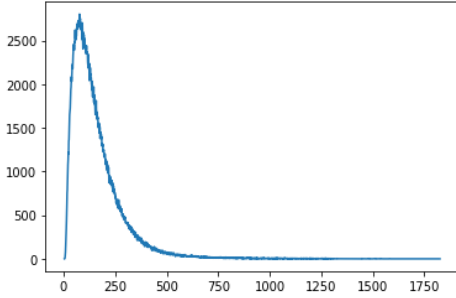

Fig. 1. Distribution of Claim Code Occurrences

Fig. 2. Distribution of Number of Encounters Per Patient

B. Problem Definition

Now we formally define the problem.

We model the problem as a classification task. A patient's claim history is composed of a sequence of encounters, denoted as $P = \{e_1, e_2, \dots\}$. Each encounter e_i is composed of a sequence of claim codes, $e_i = \{x_1, x_2, \dots\}$. Consider a list of target ADEs, and a target drug d. $y \in \{-1, 1\}$ is the classification label, where $y = 1$ indicates that drug d induced at least one ADE in the target ADE list on this patient, and otherwise $y = -1$. For a set of patients who took drug d, their claim histories before taking d along with their corresponding labels are used to train the classification model. For a target patient who has not taken drug d, the model takes his claim history so far to predict the label, i.e. whether he will experience an ADE in the target ADE list if taking d now.

Identifying ADEs from Claim History. First, we identify ADEs from claims code. Based on literature, an ADE can be identified from the claim codes by concurrent presence of selected diagnosis codes and selected indication codes [27]. Diagnosis codes are part of the ICD codes as shown in Table I. An indication code is a special type of diagnosis code indicates that a patient experienced an ADE [27], [28]. Following existing work [28], we use four categories of indication codes as shown in Table II and their corresponding ICD codes. For

example, ICD code "T46.9" represents "Other and unspecified agents primarily affecting the cardiovascular system" is an indication code, indicating an ADE related to cardiovascular system. If a diagnosis code and an indication code co-occur in an encounter, we consider an ADE occurs and the diagnosis code gives the information of the AE. For example: if a diagnosis code "I42.7" (Cardiomyopathy due to drugs and other external agents) and "T46.9" both occurs in an encounter, then "I42.7" represents an ADE.

The diagnosis codes (ICD codes) of the target ADEs and the GPI code of the target drugs are input of the problem. We consider that a target drug induces a target ADE experienced by a patient if the ICD code corresponding to the target ADE and one of the indication codes in Table II are found in the same encounter within time period N after taking the drug, but not found in the claim history before taking the drug, Specifically, as illustrated in Figure 3, suppose a patient starts to take a target drug from encounter e_{M+1}. If there is no target ADE found before encounter e_{M+1} but is recorded in encounter $e_{M'}$ along with an indication code, and the time duration between e_{M+1} and $e_{M'}$ is less than N, then we consider the target drug induces this ADE. We can also use other approaches to generate ground truth, such as human labeling.

Note that it is possible that an ADE is a result of drug-drug interaction [29]. In other words, some time multiple drugs together induce to an ADE. For any of these drug is a target drug, for this drug the corresponding claim sequence is labeled positively.

Also, N is considered the effective time of a drug to cause AE. Currently N is set to be 3 months for all the drugs. Different values of N can be used for different drugs based on the drug characteristics when the information becomes available.

TABLE II
INDICATION CODES FOR ADVERSE DRUG EVENTS

Indication Category	Description
A1	The ICD-10 code description includes the phrase 'induced by medication/drug'
A2	The ICD-10 code description includes the phrase 'induced by medication or other causes'
B1	The ICD-10 code description includes the phrase 'poisoning by medication'.
B2	The ICD-10 code description includes the phrase 'poisoning by or harmful use of medication or other causes'

IV. FIRST ATTEMPT

Since patient's medical claim history consists of a sequence of claim codes which encode medical diagnoses, procedures
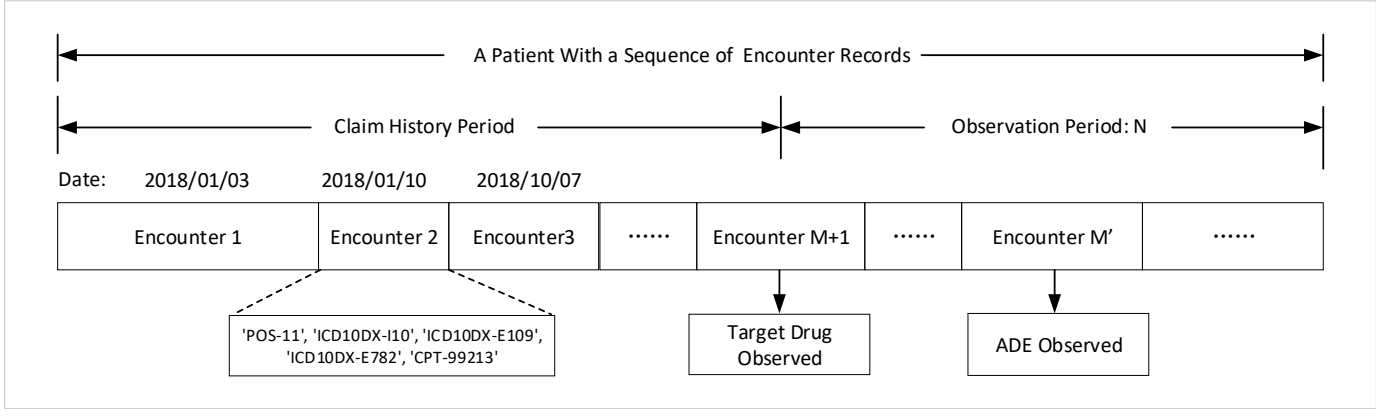
Fig. 3. Claim History Illustration

and services conducted, drugs taken and so on. Intuitively, the problem can be modeled as a sequence classification problem. Figure 4 shows a system architecture. The input is the patient's medical claim history represented as a sequence of claim codes.cThen each claim code is represented as an embedding vector. A deep learning model, Long Short-Term Memory (LSTM), is then used to learn the dependency between the claim codes in order to make the prediction whether this patient will experience a target ADE if taking a target drug.
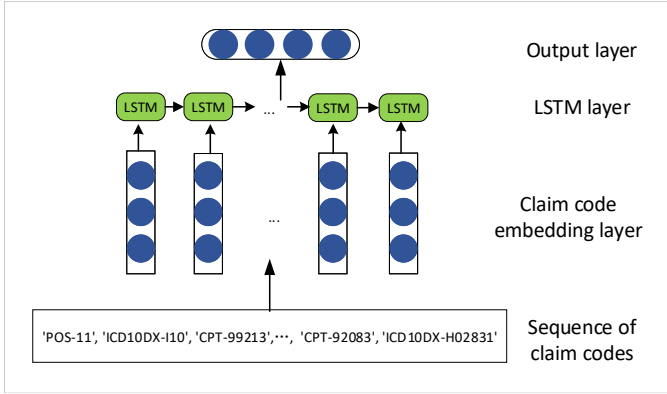

Fig. 4. The architecture of First Attempt Method

Claim Code Embedding. The claim code embedding layer generates a vector for each claim code that captures the characteristics of codes and the relationship among codes. Word embedding is widely used in deep learning based NLP techniques. Using dense and low-dimensional vectors to encode words bring computational benefits to downstream neural network model processing. Learned based on word usage, word embedding represents words that are used in similar ways using similar vectors, naturally capturing their meaning. We make the analogy that each claim code corresponds to a word, an encounter corresponds to a sentence, and a patient claims history corresponds to a document. The usage of claim code indicate their correlation, just like the usage of words in text. We use a popular word embedding method in NLP, the skip-gram model [30]. It takes as input the collection

of all patient's claim code sequences and generates a low dimensional, continuous and real-value vector for each claim code as its embedding.

Sequence Classification with LSTM. After using an embedding to represent each claim code, the sequence of claim code embedding is fed into a deep learning model to learn the claim code dependencies. We identify all patients in the training data who took a target drug. These patients' sequences of claim code embedding before taking the target drug, and the corresponding labels of whether a target ADE is observed within the L time period after taking the drug in the claim history are used to train the model. The trained model then predicts the label of each patient in the test data, based on his/her claim code embedding sequence so far.

In contrast to Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN) are designed for sequence prediction problems [1]. However, it suffers the problem of gradient vanishing or exploding [31], where gradients may grow or decay exponentially over long sequences. This makes it difficult to model long-distance correlations in claim code sequences. Recall that the average number of claim codes per patient is 1052.

We proposed to use LSTM networks instead, which are designed to overcome the vanishing gradient problem and to efficiently learn long term dependencies. LSTMs accomplish this by keeping an internal state that represents the memory cell of the LSTM neuron. This internal state controls the information flow through the cell state.

The new cell state c_j and the output h_j can be calculated as:

$$c_j = f_j \odot c_{j-1} + I_j \odot tanh(W_c[F_j, h_{j-1}] + b_c) \quad (1)$$

$$h_j = o_j \odot tanh(c_j) \quad (2)$$

where I_j, f_j and o_j denote input, forget and output gate, respectively. Finally the output layer uses a softmax function on the vector generated from the LSTM layer to make a

---

[1]The performance evaluation of CNN and and several feature-based machine learning methods are presented in Section VI.

prediction. This approach is referred as *LSTM* in the rest of paper.

V. HTNNR MODEL

After presenting the LSTM method in Section IV, now we discuss several characteristics of patient claim code history and propose a novel model named as *HTNNR Model* stands for Hierarchical Time-Aware Neural Network for ADE Risk.

A. A Hierarchical Neural Network

The LSTM method models patient claim history as a sequence of claims code. However, this approach may not accurately capture the relationship between the claims codes. Recall that the claim history actually consists of a sequence of encounters, each of which contains a sequence claim codes. There are two observations. First, the number of claim codes in different encounters can have big variation. For instance, consider three encounters illustrated in Figure 3. The first encounter represents a hospital stay, with 30 claim codes. The next encounter represents a follow-up with a specialist, with only four claim codes. The third encounter represents a visit to a primary care doctor for a flu with another four claim codes. The LSTM model ignores the encounter information, but just considers the claim code sequence where code relationships are reflected by their distances. In this example, the 1st code and the 30-th code are considered less related since their distance is 29, despite that they actually belong to the same encounter. On the other hand, the 30-th code and the 35-th one are considered as closely related since their distance is only 5. However, they actually are two encounters apart, and are not semantically closely related. The second observation is that the claims code within an encounter are actually not ordered, collectively describing an encounter event.

Based on this observation, we propose a hierarchical framework to model the input data, as shown in Figure 5. The first layer in framework generates a vector for each encounter, called *Encounter Representation*. The second layer in the framework takes the sequence of encounter vectors as input and outputs an embedded vector for each patient's claim history, referred to as *Claim History Representation*. This framework better captures the claim code relationships. Now we discuss these two layers in term.

B. Encounter Representation

The Encounter Representation takes the patient claim history as input. It has two components: a Bi-LSTM layer and a claim code attention layer. We discuss each in turn.

Bi-LSTM Representation for Encounters. Recall that the LSTM method discussed in Section IV consider claim history as a sequence of claim codes. However claim codes in an encounter do not have sequential order, but are a set of codes that collectively record an encounter event. Based on this observation, we propose to use Bi-directional Long Short Term Memory (Bi-LSTM) [32] to generate a representation of claim codes in an encounter, which are unordered. Both previous codes and following codes within an encounter are considered

by Bi-LSTM to model code dependencies. The output of the jth claim code in an encounter is calculated as:

$$h_j = \overrightarrow{h_j} \oplus \overleftarrow{h_j}, \qquad (3)$$

where $\oplus$ is an concatenation operation.

Claim Code Attention. Not all claim codes contribute equally to the semantic representation of an encounter. Attention neural networks have recently demonstrated success in document classification by learning the weights of words [33]. Hence, we apply the attention mechanism to set weights of claim codes, so that the model can focus on claim codes that are important to capture the semantics of an encounter. The encounter representation $\mathbf{v}_e$ is formed by a weighted sum of the vectors generated by Bi-LSTM.

$$E = tanh(\mathbf{H}) \qquad (4)$$

$$\alpha = softmax(w^{\mathrm{T}}E) \qquad (5)$$

$$\mathbf{v}_e = \mathbf{H}\alpha^{\mathrm{T}} \qquad (6)$$

Here $\mathbf{H}$ is a matrix consisting of vectors $[h_1, h_2, ..., h_T]$ that the Bi-LSTM layer produces, where T is the input length. w is a trained weight vector and w^{T} is a transpose.

C. Claim History Representation

Given the encounter vectors $\mathbf{v}_{e_i}$ output by the Encounter Representation layer for every encounter e_i in a claim histor, now we discuss how to generate vector for each patient's claim history.

One intuitive way is to use a LSTM model on the sequence of encounter vectors to generate a claim history vector. Indeed the sequential order of encounter indicate the temporal order of the encounter events. However, LSTM does not capture the time differences among the encounters. Referring to Figure 3. The first two encounters are 7 days apart, with the second encounter being a follow-up visit of a surgery preformed in the first encounter. The time between the second and the third encounter is 9 months, with the third encounter being a visit to a primary care doctor for a flu. As we can see from this example, two adjacent encounters that has a small time lap often refer to closely related medical issues. On the other hand, two adjacent encounters that are a long time apart likely refer to unrelated medical issues. In this case, the previous encounter has less importance to the semantics of the current encounter. Thus, sequential order itself is inadequate to capture the relationship between encounters, we should also consider the actual time differences.

We propose to use a Time-aware LSTM (TLSTM) [34] to generate a claim history vector from the sequence of encounter vectors for each patient. For each encounter, we consider not only its claims code, but also its timestamp. The major component of the TLSTM layer is the subspace decomposition applied on the memory of the previous time step. The short-term memory is adjusted proportionally to the amount of time span between two patient encounters.

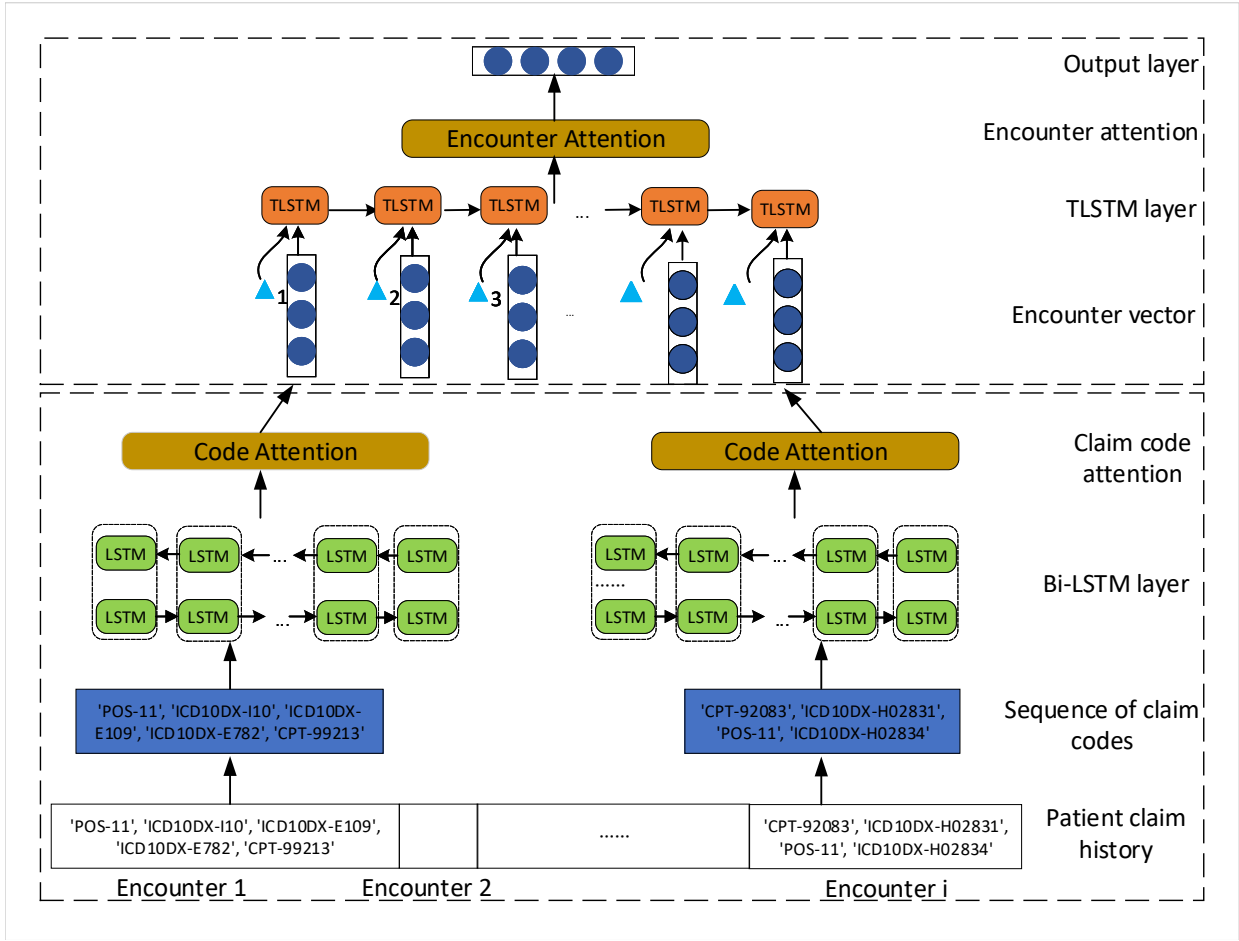$$g(\Delta_i) = 1/\Delta_i \qquad (7)$$

Fig. 5. Architecture Overview of HTNNR

$$\hat{c}_{i-1} = c_{i-1} * g(\Delta_i) \qquad (8)$$

$$c_{i-1}^* = c_{i-1}^L + \hat{c}_{i-1} \qquad (9)$$

Here Δ_i is the time span between encounter e_i and encounter e_{i-1}, c_{i-1} is short memory in LSTM, $\hat{c}_{i-1}$ is the adjusted short memory by considering time span. c_{i-1}^* is the final adjusted previous memory that combines the normal long term memory c_{i-1}^L and the adjusted short term memory. As we can see, if the gap between encounter e_i and e_{i-1} is large, which means there is no new information recorded for the patient for a long time, the dependence on the short-term memory does not play a significant role in the prediction of the current output.

In this way, the final cell state in Equation 1 is changed to:

$$c_i = f_i \odot c_{i-1}^* + I_i \odot tanh(W_c[F_i, h_{i-1}] + b_c) \qquad (10)$$

The patient claim history vectors are calculated from encounter vectors as the following:

$$h_i = TLSTM(\mathbf{v}_{e_i}, \Delta_i), i \in [1, M] \qquad (11)$$

Here $\mathbf{v}_{e_i}$ is the encounter representation for encounter e_i, M is the number of encounters before the target drug taken. Δ_i is the elapsed time between encounter e_i and e_{i-1}.

Finally the patient claim history vector is fed into an attention layer to learn the importance of different encounters to make the prediction whether the target patient will experience a target ADE.

VI. EXPERIMENT

We implemented the proposed method, referred as *HTNNR*. We have conducted extensive experiments to empirically evaluate the HTNNR model using real-life claims data. We start with discussing the model implementation, evaluation setting and comparison methods. Then we present the empirical evaluation results.

A. System Implementation

HTNNR is implemented using Python and the Hierarchical Attention model is implemented using Keras with Tensorflow backend. The experiments are run on a 20-core computer server. Existing work indicates that a large batch size may alleviate the impact of noisy data, while a small size sometimes can accelerate of convergence [35]. We varied the batch size in experiments, and set the training batch size to 256 considering the trade-off of performance and the consumption of training time and memory. To train ADE classification, we use binary

cross-entropy as the loss function. The optimizer we adopted is Adaptive Moment Estimation (Adam) which can achieve fast gradient descent [36]. We use validation-based early stopping to obtain the models that work the best with the validation data. The model with the minimum validation error are saved and used to make prediction the testing data.

B. Evaluation Setting

The data we used is provided by Inovalon. Inovalon's $MORE^2$ Registry dataset contains 500K patients. Each patient contains a sequence of encounters and each encounter contains a sequence of claim codes, with statistics presented in Section III-A.

target Drugs. We evaluated our proposed methods on 10 randomly selected drugs among all drugs, each of which has been taken by more than 20K patients in the dataset. Table III shows the GPI, description and the number of patients taking the drug.

Drug GPI code	Description	Patient Population
GPI-5818002510	Duloxetine HCl	22616
GPI-3610003000	Lisinopril	124716
GPI-4927006000	Omeprazole	138152
GPI-3400000310	Amlodipine Besylate	127326
GPI-4220003230	Fluticasone Propionate	106106
GPI-3320003010	Metoprolol Tartrate	75561
GPI-3615004020	Losartan Potassium	75570
GPI-5710001000	Alprazolam	44214
GPI-5816007010	Sertraline HCl	39258
GPI-6420001000	Acetaminophen	20618

target ADEs. ADEs are prevalent, and are not totally avoidable. The evaluation is performed on target ADEs that are severe. Table IV shows the target ADE list used in evaluation, selected based on its severity according to existing studies [37] and their occurrence in our data set. Here the occurrence means the number of the patients experienced this ADE in our data set. Other ADEs can also be used in evaluation.

Training and Testing Data. For each target drug, we extract all the patients whose claim history contains the GPI code of the drug. For each patient, we extract the claim history before taking the target drug. Then we identify the occurrence of a target ADEs within 3 months after the drug taking using the method discussed in Section III to generate the label for this instance. We split all the patients in each drug into training/testing/validation dataset with ratio 0.7/0.2/0.1. The final result is the averaged result of these 10 drugs

C. Comparison Methods

Since we are the only study that uses claims history for personalized ADE risk prediction, there is no existing work to compare. We use several baseline approaches for comparison.

- **Long Short Term Memory (LSTM):** This is the method discussed in Section IV.

TABLE IV
TARGET ADVERSE DRUG EVENTS (ADES)

ADE code (ICD 10)	Description
L29.9	Pruritis
K27.9	Stomach or intestinal ulcers
L50.9	Urticaria
T78.40	Allergic Reaction
F329	Depression
R06.00	Dyspnea
D649	Anemia
D696	Thrombocytopenia
M25.50	Arthralgia
R00.2	Palpitation
R20.2	Paresthesia
F419	Anxiety
M79.1	Myalgia
I47.2	Ventricular tachycardia
I63.0	Anorexia

TABLE V
EVALUATION OF OVERALL EFFECTIVENESS

Systems	Accuracy	Precision	Recall	AUC
Random Forest	0.78	0.65	0.21	0.75
XGBoost	0.80	0.67	0.25	0.76
LSTM	0.84	0.69	0.34	0.81
CNN	0.83	0.60	0.37	0.80
HTNNR	**0.88**	**0.84**	**0.51**	**0.89**

- **Convolutional Neural Network (CNN):** This replaces the LSTM model in the method discussed in Section IV with a CNN model. CNN has proven effectiveness in computer vision [38], natural language processing [39]
- **Random Forest:** Random forest is a classification algorithm consisting of many decisions trees [40].
- **XGBoost:** XGBoost is an implementation of gradient boosted decision tree algorithm which has been widely used in many classification tasks like emotion analysis [41] and image classification [42]

Note that every method is trained on the patients for each target drug independently. For Random Forest and XGBoost, we use Term Frequency (TF)-Inverse Document Frequency (IDF) vectors extracted from claim code sequence as features. TF-IDF has been commonly used as features in text classification tasks [43].

D. Evaluation of Overall Effectiveness

Table V shows the performance of different methods on target drugs. For each system, each number is the average performance on ten drugs in each drug group.Several observations can be made.

The proposed HTNNR method consistently achieves the best performance among these methods on all metrics. One reason is that the hierarchical attention model to differentiate the relationship of claim codes in an encounter, and the

relationship of encounters in an claim history. It further uses different neural networks, Bi-LSTM, and TLSTM, respectively, to capture their different characteristics. On the other hand, comparison systems model the input as a sequence of claims code for each patient. Furthermore, the attention layer in HTNNR gives higher weights on important claim codes and important encounters.

We also observe that the performance differences on precision and recall are much bigger than those on AUC and Accuracy. It is relatively easy for a model to perform well on AUC and Accuracy on imbalanced data. AUC represents the model overall classification ability on various thresholds. It does not reflect well the effect of minority class. Even if a method mis-classifies most or all of the minority class, its AUC value can still be high. Similarly, for imbalanced data, if a model always predicts the majority label, it will obtain a good accuracy. In our case, the target drug list has about 80% negative labels. Thus most methods perform similarly on AUC and Accuracy. High AUC and Accuracy can be misleading in some imbalanced data. On the other hand, achieve high precision and recall are much more challenging. In the following, we focus the analysis on precision and recall.

E. Evaluation on single drug

Table V shows the average results on the 10 drugs. Now we zoom in to a single drug. We randomly select a drug from target durg list, GPI-3320003010, and evaluate the performance of comparison systems, and HTNNR on its ADE risk prediction, as shown in Figure 6. Here we only show the precision and recall, as the performance differences of Accuracy and AUC are similar as the result represented in Table V. There are several things worth mention. First, the HTNNR model performs better than the comparison systems, consistent with the evaluation shown in Table V.

we also observe the improvement on recall is higher than that on precision. Hierarchical framework helps to find more shared ADE characteristics among the drugs. At the same time, more noisy information is introduced. Thus, the recall benefits more from training data from multiple drugs than precision.
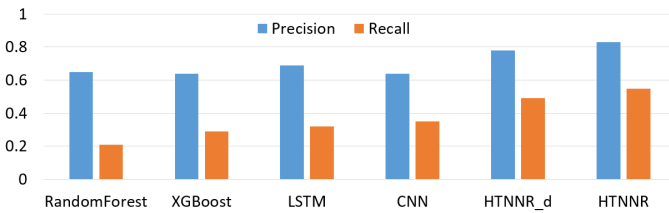

Fig. 6. GPI-3320003010 (# of Patients Taken: 75,561)

F. Evaluation of Patient Claim History Length

All the results shown so far takes as input each patient's entire medical claim history before taking a target drug to train each model. To evaluate how the patient claim history impact the personalized ADE prediction, we evaluated the performance on different time length of medical history considered.

Figure 7 shows the performance vary with varying length of each patient's medical history used to train HTNNR. The medical history always ending at the time when a target drug is recorded in the claim history, with duration count backward. The results show that using 3 month of claim history generates better performance than using 1 month of claim history, since the model can benefit from a larger dataset. After 3 months, the longer history considered, the better recall, and the worse precision. The reason is that longer history data can help the model to find more characteristics of patients and potential drug interactions, but at the same time, introduce more noisy information. In a real application, we can adjust the history length to be considered depends on which metrics is more important in the application.
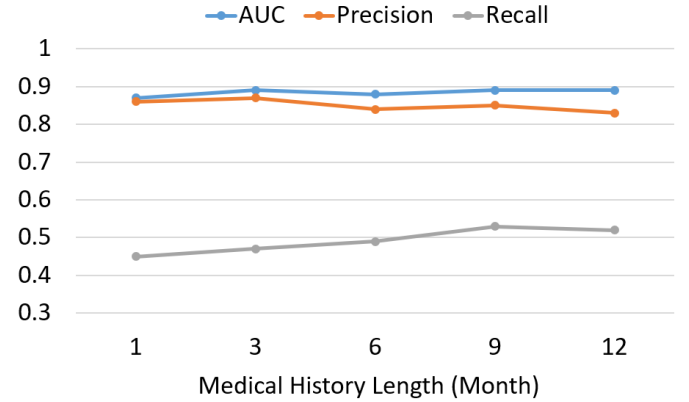

Fig. 7. Evaluation on Different Length of Claim History

To summarize, HTNNR achieves the best effectiveness in all evaluation metrics among all methods tested.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we studied how to use patient claims history for personalized ADE risk prediction. We propose the HTNNR model that captures the characteristics of claim codes and their relationship. It has a hierarchical framework. The first layer first generates embedding for claim codes, and then generate a vector for each encounter using a Bi-LSTM model with an attention layer. The second layer takes the sequences of encounter vectors as input and uses a time-aware neural network model to generate claim history representation that capture the temporal order of encounters. The empirical evaluation show that the proposed HTNNR model is effective and efficient, especially for rare drugs.

Since claim history is updated on daily basis, as future work we will investigate how to incrementally train the model based on the new information available without re-training the model from scratch every time.

REFERENCES

[1] J. R. Nebeker, P. Barach, and M. H. Samore, "Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting," *Annals of internal medicine*, vol. 140, no. 10, pp. 795–801, 2004.

[2] K. B. Sonawane, N. Cheng, and R. A. Hansen, "Serious adverse drug events reported to the fda: analysis of the fda adverse event reporting system 2006-2014 database," *Journal of managed care & specialty pharmacy*, vol. 24, no. 7, pp. 682–690, 2018.

[3] , "Adverse drug events," 2018. [Online]. Available: https://www.fda.gov/drugs/drug-interactions-labeling/preventable-adverse-drug-reactions-focus-drug-interactions

[4] health.gov, "Adverse drug events," 2020. [Online]. Available: https://health.gov/our-work/health-care-quality/adverse-drug-events

[5] L. Hazell and S. A. Shakir, "Under-reporting of adverse drug reactions," *Drug safety*, vol. 29, no. 5, pp. 385–396, 2006.

[6] J.-L. Montastruc, A. Sommet, H. Bagheri, and M. Lapeyre-Mestre, "Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database," *British journal of clinical pharmacology*, vol. 72, no. 6, p. 905, 2011.

[7] S. J. Evans, P. C. Waller, and S. Davis, "Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports," *Pharmacoepidemiology and drug safety*, vol. 10, no. 6, pp. 483–486, 2001.

[8] P. LePendu, S. V. Iyer, A. Bauer-Mehren, R. Harpaz, J. M. Mortensen, T. Podchiyska, T. A. Ferris, and N. H. Shah, "Pharmacovigilance using clinical notes," *Clinical pharmacology & therapeutics*, vol. 93, no. 6, pp. 547–555, 2013.

[9] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *Journal of biomedical informatics*, vol. 57, pp. 333–349, 2015.

[10] G. Wang, K. Jung, R. Winnenburg, and N. H. Shah, "A method for systematic discovery of adverse drug events from clinical notes," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1196–1204, 2015.

[11] K. Haerian, D. Varn, S. Vaidya, L. Ena, H. Chase, and C. Friedman, "Detection of pharmacovigilance-related adverse events using electronic health records and automated methods," *Clinical Pharmacology & Therapeutics*, vol. 92, no. 2, pp. 228–234, 2012.

[12] C. McMaster, D. Liew, C. Keith, P. Aminian, and A. Frauman, "A machine-learning algorithm to optimise automated adverse drug reaction detection from clinical coding," *Drug safety*, vol. 42, no. 6, pp. 721–725, 2019.

[13] J. M. Bos, G. A. Kalkman, H. Groenewoud, P. M. van den Bemt, P. A. De Smet, J. E. Nagtegaal, A. Wieringa, G. J. van der Wilt, and C. Kramers, "Prediction of clinically relevant adverse drug events in surgical patients," *PloS one*, vol. 13, no. 8, p. e0201645, 2018.

[14] R. Liu, M. D. M. AbdulHameed, K. Kumar, X. Yu, A. Wallqvist, and J. Reifman, "Data-driven prediction of adverse drug reactions induced by drug-drug interactions," *BMC Pharmacology and Toxicology*, vol. 18, no. 1, p. 44, 2017.

[15] G. Jiang, H. Liu, H. R. Solbrig, and C. G. Chute, "Mining severe drug-drug interaction adverse events using semantic web technologies: a case study," *BioData mining*, vol. 8, no. 1, p. 12, 2015.

[16] J. D. Stein, F. Lum, P. P. Lee, W. L. Rich III, and A. L. Coleman, "Use of health care claims data to study patients with ophthalmologic conditions," *Ophthalmology*, vol. 121, no. 5, pp. 1134–1141, 2014.

[17] C. Wang, X.-J. Guo, J.-F. Xu, C. Wu, Y.-L. Sun, X.-F. Ye, W. Qian, X.-Q. Ma, W.-M. Du, and J. He, "Exploration of the association rules mining technique for the signal detection of adverse drug events in spontaneous reporting systems," *PloS one*, vol. 7, no. 7, p. e40561, 2012.

[18] J. M. Reps, U. Aickelin, J. Ma, and Y. Zhang, "Refining adverse drug reactions using association rule mining for electronic healthcare data," in *2014 IEEE International Conference on Data Mining Workshop*. IEEE, 2014, pp. 763–770.

[19] H. Z. Lo, W. Ding, and Z. Nazeri, "Mining adverse drug reactions from electronic health records," in *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013, pp. 1137–1140.

[20] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman, "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 413–419, 2013.

[21] P. Dias, A. Penedones, C. Alves, C. F Ribeiro, and F. B Marques, "The role of disproportionality analysis of pharmacovigilance databases in safety regulatory actions: a systematic review," *Current drug safety*, vol. 10, no. 3, pp. 234–250, 2015.

[22] A. B. Chapman, K. S. Peterson, P. R. Alba, S. L. DuVall, and O. V. Patterson, "Detecting adverse drug events with rapidly trained classification models," *Drug safety*, vol. 42, no. 1, pp. 147–156, 2019.

[23] B. Dandala, V. Joopudi, and M. Devarakonda, "Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks," *Drug safety*, vol. 42, no. 1, pp. 135–146, 2019.

[24] H. Kwak, M. Lee, S. Yoon, J. Chang, S. Park, and K. Jung, "Drug-disease graph: Predicting adverse drug reaction signals via graph neural network with clinical data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 633–644.

[25] T. Islam, N. Hussain, S. Islam, and A. Chakrabarty, "Detecting adverse drug reaction with data mining and predicting its severity with machine learning," in *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2018, pp. 1–5.

[26] A. Valeanu, C. Damian, C. D. Marineci, and S. Negres, "The development of a scoring and ranking strategy for a patient-tailored adverse drug reaction prediction in polypharmacy," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[27] S. R. Walter, R. O. Day, B. Gallego, and J. I. Westbrook, "The impact of serious adverse drug reactions: a population-based study of a decade of hospital admissions in new south wales, australia," *British journal of clinical pharmacology*, vol. 83, no. 2, pp. 416–426, 2017.

[28] C. M. Hohl, A. Karpov, L. Reddekopp, and J. Stausberg, "Icd-10 codes used to identify adverse drug events in administrative data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 547–557, 2014.

[29] M. M. Alvim, L. A. da Silva, I. C. G. Leite, and M. S. Silvério, "Adverse events caused by potential drug-drug interactions in an intensive care unit of a teaching hospital," *Revista Brasileira de terapia intensiva*, vol. 27, no. 4, p. 353, 2015.

[30] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling." in *LREC*, vol. 6, 2006, pp. 1222–1225.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[33] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[34] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 65–74.

[35] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry, "Augment your batch: better training with larger batches," *arXiv preprint arXiv:1901.09335*, 2019.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] A. Gottlieb, R. Hoehndorf, M. Dumontier, and R. B. Altman, "Ranking adverse drug reactions with crowdsourcing," *Journal of medical Internet research*, vol. 17, no. 3, p. e80, 2015.

[38] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synthesis Lectures on Computer Vision*, vol. 8, no. 1, pp. 1–207, 2018.

[39] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[40] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[41] M. Jabreel and A. Moreno, "Eitaka at semeval-2018 task 1: An ensemble of n-channels convnet and xgboost regressors for emotion analysis of tweets," *arXiv preprint arXiv:1802.09233*, 2018.

[42] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, "A novel image classification method with cnn-xgboost model," in *International Workshop on Digital Watermarking*. Springer, 2017, pp. 378–390.

[43] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf*idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.