

Prediction of drug adverse events using deep learning in pharmaceutical discovery

Chun Yen Lee and Yi-Ping Phoebe Chen

Corresponding author: Yi-Ping Phoebe Chen, Tel.: 61-3-94796768. E-mail: phoebe.chen@latrobe.edu.au

Abstract

Traditional machine learning methods used to detect the side effects of drugs pose significant challenges as feature engineering processes are labor-intensive, expert-dependent, time-consuming and cost-ineffective. Moreover, these methods only focus on detecting the association between drugs and their side effects or classifying drug–drug interaction. Motivated by technological advancements and the availability of big data, we provide a review on the detection and classification of side effects using deep learning approaches. It is shown that the effective integration of heterogeneous, multidimensional drug data sources, together with the innovative deployment of deep learning approaches, helps reduce or prevent the occurrence of adverse drug reactions (ADRs). Deep learning approaches can also be exploited to find replacements for drugs which have side effects or help to diversify the utilization of drugs through drug repurposing.

Key words: deep learning; pharmacovigilance; adverse drug reactions

Introduction

Predicting the side effects of drugs is important during the drug discovery process. Detecting and predicting adverse drug reactions (ADRs) during post-marketing surveillance is equally important as ADRs are the leading cause of mortality and morbidity. Statistics show that each year, between 44,000 and 98,000 deaths occur due to medical errors and 7000 deaths occur due to the occurrence of ADRs caused by medicine taken at a recommended dosage [7]. Very often, ADRs with serious adverse consequences such as those caused by rosiglitazone maleate are reported only after their introduction in the market [2].

A study conducted by Pierre et al. revealed that from 1976 to 2010, 34 drugs were withdrawn from the United States due to ADRs. Onakpoya et al. reported that 462 medicinal products were withdrawn from the United States between years 1953 and 2014 and 43 drugs were withdrawn worldwide [35]. According to the Food and Drug Administration (FDA) website, 113 drugs

that were approved from 2015 to 2017 were withdrawn from the market after 2017. For example, Vioxx, approved by the FDA in 1999, was prescribed to over 80 million people worldwide to treat arthritis patients with chronic pain, but was withdrawn 5 years later due to an enhanced risk of heart attack and stroke [32, 33].

Many patients are on multiple prescriptions and over-the-counter medications in real-life settings. It was estimated that 82% of Americans consume at least one drug and 29% consume five or more drugs [37]. Drug–drug interactions (DDIs) due to the simultaneous use of two or more drugs are another significant problem for drug safety, accounting for up to 30% of unexpected ADRs [37].

Liu et al. defined a synergistic DDI-induced ADR as arising from simultaneous interactions between co-administered drugs and their protein targets, which caused a new or enhanced ADR beyond what either drug could trigger on its own [25].

Chun Yen Lee is a PhD student at the Department of Computer Science and Information Technology, La Trobe University. Her research interests include the application of artificial intelligence in pharmacovigilance.

Yi-Ping Phoebe Chen is a professor of the Department of Computer Science and Information Technology, La Trobe University. Her research interests include bioinformatics, multimedia technologies and knowledge discovery.

Submitted: 10 November 2019; Received (in revised form): 8 February 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

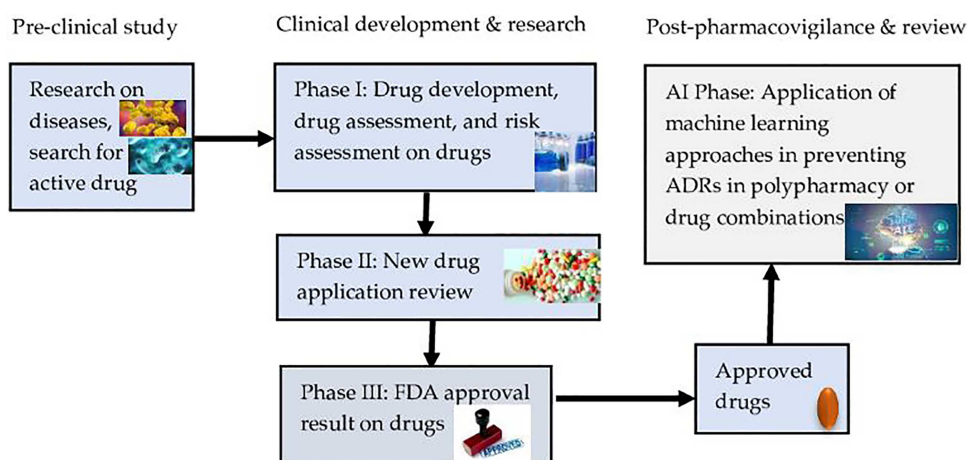


Figure 1. The pharmacovigilance process comprises three stages (preclinical study, clinical development and research, post-pharmacovigilance and review). In the preclinical study stage, a drug undergoes extensive laboratory testing where a drug in testing is experimented with animals to study its potential properties, safety and its value as a new therapy. The results from the preclinical phase are then submitted to the regulatory authorities for approval to be tested on people in clinical trials. There are three phases in the clinical development and research stage. All new drugs have to go through these phases before they can be prescribed to patients. Phase I is about developing the drug and assessing the effectiveness and safety of a drug. It is mainly about finding out whether the drug is safe for human use. Phase II aims to assess the effectiveness of a drug at treating a particular disease. It also aims at finding the best method for drug delivery, e.g. in the form of tablets, sprays, injections, etc. Many treatments do not make it to Phase II. Phase III is the FDA approval result on the drugs. In the post-pharmacovigilance and review stage is the AI Phase, where machine learning techniques are used to reduce or prevent any possible ADRs in polypharmacy or drug combinations.

Generally, when two drugs interact in a human body, each drug will cause either therapeutic effects or adverse reactions or enhanced adverse effects. According to Vilar et al., in some extreme cases, adverse reactions caused by DDIs could lead to death. In June 2001, the drug cerivastatin caused 31 cases of fatal rhabdomyolysis, 12 of which involved the concomitant use of cerivastatin and gemfibrozil [45].

There are two major stages in the detection of ADRs: pre-marketing review and post-marketing surveillance as shown in Figure 1. Pre-marketing review which focuses on identifying the risk associated with drugs is required before any new drug is approved by the FDA for marketing. The risks are then established and clearly communicated to prescribers and consumers. A preclinical study generally focuses on the detection of ADRs induced by a single drug and rarely investigates a DDI occurrence [16]. Clinical trials to detect a DDI are usually carried out in selective patient populations, with a relatively small number of patients, and involve only a short follow-up period [16]. Therefore, DDI detection depends heavily on post-marketing surveillance which involves the systematic detection and evaluation of drugs after they are marketed.

Most of the extraction methods for ADRs in the post-marketing period were based on traditional statistical and machine learning methods which are time-consuming, cost-ineffective and expert-dependent annotation processes. Recently, deep learning has demonstrated its superiority over time-consuming and costly hand-crafted feature-based traditional machine learning models [21, 38, 44, 45]. Because of their hierarchical learning structures and the focus of these methods on representation learning, deep architectures have the potential to integrate diverse datasets across heterogeneous data types and provide greater generalization [28]. In view of the availability of a huge amount of data and the capability of deep learning to handle heterogeneous, complex and multimodality data from diverse sources, coupled with its superior feature learning ability, in this paper, we carry out (a) a review on the application of deep learning approaches in extracting ADRs caused by single drugs and drug-drug combinations. To the best of our knowledge, this

paper is the first to conduct an overall review on studies related to the detection and classification of ADRs using deep learning approaches; (b) we propose to integrate multidimensional drug properties across heterogeneous data sources to detect and predict ADRs induced by DDIs using deep learning approaches; (c) we identify the challenges faced in using deep learning and suggest several emerging factors that can help to enhance the accuracy and robustness of deep learning models for ADR prediction.

Deep learning architecture

Essentially, there are different types of deep learning (DL) models which include convolutional neural networks (CNNs), recurrent neural networks (RNNs) and its variants long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional LSTM (BiLSTM) and many others. A brief description of these mainstream architectures is introduced as follows:

Convolutional neural networks

A convolutional neural network (CNN) is a feedforward neural network (FNN) with three types of layers, namely, the convolution layer, pooling layer and fully connected layer [18]. The same as the architecture of FNN, there is no connection between hidden nodes in the same layers, but there is a connection between nodes in adjacent layers. Different types of layers are used for different types of input data modality. For sequence signals such as language, layers can be formed with a one-dimensional array; for images or audio, layers formed with two-dimensional arrays can be applied; while for video, layers can be formed with three-dimensional arrays [18]. CNN is capable of capturing global and local features, and it performs well in processing data with a grid-like topology.

Recurrent neural network and its variants

A recurrent neural network (RNN) is another DL network. It is suitable for modeling sequential data. Even though RNN is

powerful in capturing long-distance dependencies, it was later found that ordinary RNN suffers from gradient vanishing/exploding problems [28, 38]. To resolve this problem, popular RNN variants, long short-term memory (LSTM) and gated recurrent unit (GRU) models, have been introduced. The LSTM architecture contains an input gate, forget gate and output gate, together with a recurrent cell. The input to the LSTM cell is multiplied by the activation of the input gate, while the previous values are multiplied by the forget gate. The network only interacts with the LSTM cell via the gates [28, 38]. GRU, another variant of RNN, simplifies the LSTM architecture by combining the forget and input gates into a single update gate and merges the cell state with the hidden state [38].

Another inherent issue of unidirectional RNN is its dependency on the previous context. A complete dependency on the previous context can lead to less accurate prediction. Bidirectional LSTM (BiLSTM), which is another important modification of basic RNN architecture, is used to address this issue [38]. Bidirectional LSTMs contain two chains of LSTM cell flows in both forward and backward directions, incorporating information from the preceding as well as the future context of the current term to some extent. It can learn the high-level syntactic meaning of a whole sentence and pass outputs to the next layer. Hence, it can better capture the global semantic representation of an input [38].

In recent years, the concept of attention has been introduced to neural network to equip the DL network with the ability to (a) flexibly select the latent representation of the input data, (b) calculate attention signals based on the latent representation and (c) assign attention signals to generate the weighted latent representation of the input data [6]. The attention mechanism is applied to input word embeddings to improve the ability of the BiLSTM to identify the words of a sentence which are the most influential for determining the relationship between two drug candidates [6]. It was reported that LSTM with attention input (Att-BiLSTM) is distinctly more superior than BiLSTM which has no attention mechanism. This is because it provides more targeted semantic matching and the model is able to explicitly find important clue words. It can overcome the bias deficiency of BiLSTM which omits some important previous information when processing long sentences. To get the optimal fixed length features, max pooling is then applied to the output of BiLSTM. These features are then fed to fully connect neural layer, followed by softmax layer to obtain final classification [6].

To further improve the performance of the BiLSTM model with single input attention, in addition to using network architecture with position-aware attention (PM-LSTM), multiple attention layers are added to the BiLSTM model [54]. The attention layers are added to both the word level and sentence level. The word-level attention can automatically learn the weight of each word and hence can determine the importance of the individual encoded words in every row. It transforms the sentence matrix to vector representation. The sentence-level attention then generates the final representation by combining several relevant sentences which have the same drug pairs. This helps to contribute to the understanding of drug-drug interactions from all the sentences. The softmax layer is then applied to classify the drug pairs. The BiLSTM with a multiple attention mechanism is reported to outperform models with single attention or with no attention [54].

In recent years, word embedding has successfully been applied in natural language processing (NLP) tasks such as sequence labeling. The concept behind word embedding is that it is a learned representation for text where words have the same

meaning and carry similar representation and semantically similar words usually have close embedding vectors [30]. For common word embedding tools such as word2vec, it learns low-dimensional continuous vector representation of words from an unlabeled large text corpus, based on the word's context in different sentences [30]. Word embedding has shown improvement compared to traditional feature-based models because it does not rely heavily on hand-crafted features [30].

Main data sources used for the detection and prediction of ADRS and DDI

Diverse data sources have been used by researchers for post-marketing drug safety surveillance. Some of the primary traditional data sources used by researchers include spontaneous reporting systems (SRS) (which cover the spontaneous reporting of adverse drug events by healthcare professionals or patients), clinical narratives written by healthcare professionals and electronic health records that store records of diagnoses. Other useful big data sources include social media posts such as tweets, blogs and forums related to health. A vast amount of big data for biomedical research are extracted from DrugBank, Side Effect Resource (SIDER) and others.

The DrugBank database contains the biological, chemical and phenotypic information of drugs. The database combines detailed chemical, pharmacological and pharmaceutical data of drugs together with comprehensive drug target information including drug sequence, structure and pathway information [48].

The SIDER database contains information on approved drugs and their reported side effects, extracted from public documents and package inserts [20]. The OFF-SIDES database contains the off-label side effects of drugs [42] which are generated using the FDA Adverse Event Reporting Systems (FAERS) which collects reports from doctors, patients and drug companies. TWOSIDES contains information on the side effects caused by combinations of drugs [42]. Similarly, the TWOSIDES dataset is generated from adverse event reporting systems.

The PubChem Compound database provides drug structure information, while the KEGG database, which is a comprehensive database for approved drugs marketed in Japan, the United States and Europe, stores information on chemical structures, targets, metabolizing enzymes and other drug features [19]. It also stores information on protein pathways. The relationship between protein and drugs is obtained from the STITCH (Search Tool for Interactions of Chemicals) database, which integrates various chemical and protein networks [40].

The DDI corpus was developed for the DDI Extraction 2013 challenge (<http://www.cs.york.ac.uk/semeval-2013/task9/>), with the goal of providing a common framework for the evaluation of extraction techniques applied for the recognition of pharmacological substances and the detection of DDIs from biomedical texts. It is a gold standard corpus annotated with pharmacological substances as well as the interactions between them. It is the first corpus to include pharmacodynamics (PD) to show the pharmacological effects of one drug modified by the presence of another drug. It also includes pharmacokinetic (PK) DDIs to show the result from the interference of drug absorption, distribution, metabolism and/or elimination of a drug by another drug [13].

Drugs, when used for the treatment of diseases or illness, may introduce perturbations to our biological system which consists of various molecular interactions. Some ADRs can be detected by analyzing the biological properties of drugs, while

others require consideration of chemical properties. Generally, both the biological and chemical features of drugs must be considered simultaneously for the more accurate detection and classification of side effects. Drug information can be harvested from the DrugBank, SIDER and STITCH databases, and drug molecular structures can be downloaded from PubChem for the detection and classification of ADRs. For DDI detection and classification, the DDI corpus developed for the DDI Extraction 2013 challenge is widely used. On the other hand, for the prediction of ADRs and DDI, molecular fingerprints such as PubChem fingerprints and SMILES strings are utilized to explore the association between chemical substructures with ADRs induced by single drugs or DDI.

Review of the detection of ADRs caused by single drugs using deep learning

ADR detection is treated as a sequence labeling problem. Traditional approaches, such as state-of-the-art conditional random fields have been used to perform sequence labeling tasks where the learning methods account for the surrounding context of a given input word [3, 7, 24]. However, traditional methods such as conditional random fields (CRF) suffer from limitations because they are not designed to learn from the dependencies which lie in the surrounding context. The model can only consider the target word and its neighboring words within a fixed-width window. If the context window is too small, it is unable to include all the information. On the other hand, too large a context window will compress irrelevant words and vital information together, and this may lead to overfitting [11]. The weakness of statistical models and traditional machine methods which require laborious and time-consuming hand-crafted feature engineering processes has motivated researchers to turn to deep learning for a better solution [29, 31, 41].

Jagannatha and Hong used LSTM and GRU frameworks to extract medical events and their attributes from unstructured text in electronic health records. They trained the RNN frameworks on sentence and document levels [17]. The sentence-level neural networks were fed with one sentence at a time. As shown in Table 1, Jagannatha and Hong consistently demonstrated that all RNN models significantly outperformed the baseline CFR model [17]. Of all the RNN models, the GRU model exhibited an F-score of at least one percentage point higher than the rest of the models. These models were able to remember information across a different range of dependencies. On the other hand, CFR models with hand-crafted features which used a bag-of-words representation lost a lot of information in the process due to the need to use fixed context windows [17]. Table 1 gives a summary of the crucial characteristics of some studies on the detection and extraction of ADRs using deep learning approaches.

Huynh et al. focused on using different neural network (NN) architectures for ADR classification (Table 1) [15]. They proposed two new neural network models, namely, convolutional recurrent neural networks (CRNNs), where convolutional neural networks were concatenated with recurrent neural networks, and convolutional neural networks with attention (CNNA), where attention weights were added to convolutional neural networks. They evaluated the neural network (NN) architectures on a Twitter dataset which contained informal language, and they also tested their NN architectures on an adverse drug event (ADE) dataset constructed using sampling from MEDLINE case reports. Their experiment results showed that all their neural network (NN) architectures outperformed

the traditional maximum entropy classifiers. Similarly, all their NN architectures performed better on the Twitter dataset. CNNA allowed the visualization of the attention weights of words when making classification decisions; hence it was appropriate for the extraction of words describing ADRs.

Cocos et al. developed a bidirectional long short-term memory (BiLSTM RNN) network-based model to process text in social media posts as a sequence of words [7]. They evaluated three BiLSTM variants namely:

- Method 1 (BiLSTM-M1): Word embedding values randomly initialized and treated as learnable parameters.
- Method 2 (BiLSTM-M2): Word embedding values initialized with a publicly available pre-trained database and treated as learnable model parameters.
- Method 3 (BiLSTM-M3): Word embedding values initialized as in method 2 but treated as fixed constants.

They passed word embedding-based features to BiLSTM models for training and generating sequence labels instead of using human-engineered features. Their BiLSTM models were able to learn dependencies more efficiently on all previous outputs as well as current input over longer intervals. Moreover, these models were able to process sequences in forward and backward directions, enabling them to learn dependencies in both directions. Of the three methods, the BiLSTM-M3 model significantly outperformed all the other models in terms of F-measure. Both BiLSTM models initialized with pre-trained embeddings (BiLSTM-M2 and BiLSTM-M3) performed significantly better than the BiLSTM-M1 model with randomly initialized embeddings because pre-trained word embedding is more effective in capturing the semantic similarity of words.

Luo proposed an LSTM RNN model to classify the relations from clinical notes using the i2b2/VA relation classification challenge dataset (Table 1) [28]. He showed that his LSTM RNN model with only a word embedding feature and no manual feature engineering achieved results which were comparable to the state-of-the-art systems on the i2b2/VA relation classification challenge. The results support the use of the word-embedded LSTM RNN model, which does not require manual feature engineering to classify the relations between medical concepts. In addition to this finding, he also showed that the LSTM RNN model which was word-embedded with clinical MIMIC-III corpus outperformed the LSTM models with general domain embedding (Table 1). Hence, we can conclude that a model with word embedding in the medical related domain is preferable to a hand-crafted engineering model or a model with general domain embedding.

In view of the shortcoming of supervised learning methods which rely heavily on large annotated datasets, Gupta et al. proposed a novel semi-supervised sequence labeling method based on LSTM RNN to capture long-term dependencies (Table 1) [12]. The drug names used as keywords for searching related tweets are humira, dronedarone, lamictal, pradaxa, paxil, zoledronic acid, trazodone, enbrel, cymbalta and quetiapine. They used a semi-supervised LSTM RNN model to predict a drug name with the drugs' context extracted from tweets. To avoid mapping the input drug name to the output drug name, they masked all the drug names in the input with a single token. Instead of using hand-crafted features, word embedding-based features were passed to a BiLSTM RNN model. They demonstrated that by leveraging a large unlabeled corpus on social media, their method outperformed the fully supervised learning baseline which relies heavily on large manually annotated corpus. It is also shown clearly in Table 1 that the model which used word

Table 1. Summary of the performance of deep learning models used for detection of ADRs from single drugs under different settings

Study	Data source	Purpose	Techniques	Model	Recall	Precision	F-score
Jagannatha et al. [16]	Unstructured text in electronic health record	Medication, diagnosis and adverse drug event detection	Used LSTM and GRU framework for information extraction in the unstructured text in electronic health record	Cross validated micro-average of precision, recall and F-score for all medical tags			
				CRF-noncontext	0.6562	0.7332	0.6925
				CRF-context	0.6806	0.7711	0.7230
				LSTM-sentence	0.8024	0.7803	0.7912
				GRU-sentence	0.8013	0.7802	0.7906
				LSTM-document	0.8050	0.7796	0.7921
				GRU-document	0.8126	0.7938	0.8031
Huynh T. et al. [14]	Two datasets were used—Twitter dataset and MEDLINE case reports	Adverse drug reaction classification	Two methods were used: Convolutional recurrent neural network (CRNN) that stacks a convolutional layer on top of a recurrent layer; convolutional neural network with attention (CNNA) that added one-filter convolutional layer on top of the direct outputs from the first convolutional layer	Adverse drug reaction classification results on the Twitter			
				CNN	0.57	0.47	0.51
				CRNN	0.55	0.49	0.51
				RCNN	0.59	0.43	0.49
				CNNA	0.66	0.40	0.49
				Adverse drug reaction classification results on the ADE datasets			
				CNN	0.89	0.85	0.87
				CRNN	0.86	0.82	0.84
				RCNN	0.89	0.81	0.83
				CNNA	0.84	0.82	0.83
Cocos A. et al. [8]	Twitter dataset	ADR detection	Developed bidirectional LSTM model that performed sequence labeling tasks. The input features were word embedding vectors	F-measure, precision and recall achieved by each model over 10 training and evaluation rounds			
				BiLSTM-M1	0.6332	0.6457	0.6272
				BiLSTM-M2	0.8070	0.6047	0.6858
				BiLSTM-M3	0.8286	0.7043	0.7549
Luo Yuan [25]	Clinical notes on Medical Information for Intensive Care database	Classified relations from clinical notes	Used word embedding to capture semantic similarity of words. Pre-train word vector on clinical notes using word2vec. Used LSTM to capture sequential patterns of data and classify relations from clinical notes	Word embedding trained on Google News corpus			
				Segment LSTM	0.629	0.665	0.647
				Sentence LSTM	0.596	0.662	0.628

Continued

Table 1. Continue

Study	Data source	Purpose	Techniques	Model	Recall	Precision	F-score
Gupta S.et al. [11]	Twitter dataset collected during the period of 2007–2010	Extraction of ADRs mention using 81 drug names as keyword search terms	Used semi-supervised sequence labeling method based on bidirectional LSTM model that relied on a small labeled data and a large unlabeled data for training. Use Keras for implementation	Word embedding trained on medical corpus			
				Segment LSTM	0.636	0.674	0.655
				Sentence LSTM	0.632	0.650	0.641
				Semi-supervised BiLSTM (SS-BiLSTM) under different word embedding settings and different unlabeled settings			
				SS-BiLSTM (with drug mask removed)	0.780	0.723	0.747
				SS-BiLSTM (with labeled tweets dictionary only)	0.769	0.727	0.745
				SS-BiLSTM (with Google News vectors)	0.774	0.708	0.736
				SS-BiLSTM (with medical embeddings)	0.716	0.642	0.673
				ADE classification performance and comparisons for different models			
Lee K et al. [18]	Twitter dataset	Automatic classification of ADE tweets	Used semi-supervised convolutional neural network (CNN)	Model	0.5422	0.6429	0.5882
				1 T-Random			
				Model 2	0.5843	0.5879	0.5861
				Sent-Health			
				Model 3 T-Drug	0.5181	0.6324	0.5695
				Model	0.5843	0.6467	0.6139
				4 T-Health-Condition			
				Model	0.5663	0.6573	0.6084
				5 T-Health-Condition*			
				Model 6 T-Drug-Condition-Sent-Health	0.6024	0.6711	0.6349
				Model 7 T-Drug-Condition*-Sent-Health	0.6084	0.6733	0.6392
				Model	0.5422	0.6429	0.5882
				1 T-Random			

Continued

Table 1. Continue

Study	Data source	Purpose	Techniques	Model	Recall	Precision	F-score
Tutubalina E et al. [36]	Annotated CADEC corpus which consisted of 1250 medical forum posts taken from Askapatient.com	Model the sequence of labels for ADR extraction	Employed CNN to extract character-level features instead of deploying hand-crafted features. Introduced a supervised joint model which combined CRF with bidirectional LSTM for extraction of ADRs	Comparison results of different types of models			
				1-layer LSTM	0.6587	0.5798	0.6167
				2-layer LSTM	0.7044	0.6362	0.6686
				3-Layer LSTM	0.7022	0.6588	0.6798
				4-layer LSTM	0.7093	0.6689	0.6885
				1-layer GRU	0.6772	0.5862	0.6284
				2-layer GRU	0.7093	0.6384	0.6720
				3-layer GRU	0.7191	0.6675	0.6923
				4-layer GRU	0.7262	0.6565	0.6896
				2-layer LSTM+CRF	0.6973	0.6947	0.6960
				2-layer LSTM+CNN + CRF	0.7039	0.5798	0.6922
				3-layer LSTM+CNN + CRF	0.7066	0.6362	0.6965
				LSTM+CNN + CRF			

^aRecall calculates how many of the actual positives the model captures through labeling it as positive.

^bPrecise measures the number of actual positive cases out of those predicted positive cases.

^cF1 score is used to seek a balance between precision and recall due to uneven data distribution.

embeddings trained on a large domain-agnostic Twitter corpus outperformed the system using word embeddings trained on the Google News corpus only. The former performed better than the medical domain-specific word embeddings. This could be due to the fact that the extraction of ADRs from social media requires more language structure and semantic information.

The challenges faced when using social platforms for the real-time detection of ADRs motivated Lee et al. to explore the use of a semi-supervised convolutional neural network (CNN) for the automatic classification of ADE in tweets [22]. They built several semi-supervised CNN models with different types of unlabeled data. In total, seven models were created. The first model, Model-1 T-Random, used a massive collection of random tweets for word embeddings. The second model, Model-2 Sent-Health, used a corpus of health-related texts from the medical concept to lay term dictionaries, medical concept terms for social media, biomedical literature and UMLS medical concept definition. The third model (Model-3 T-Drug) was trained using tweets with drug names as unlabeled data, while the fourth model (Model-4 T-Health-Condition) was built using tweets that mentioned health conditions as unlabeled data. The fifth model (Model-5 T-Health-Condition) was trained with a large text corpus created by combining random tweets with health-related texts. For the sixth model (Model-6 T-Drug-Condition-Sent-Health), they combined the tweets for T-drug, T-health-condition and sentences from Sent-Health. The last model, Model-7 T-Drug-Condition-Sent-Health, combined multiple data sets similar to training Model 6 but replaced the tweets from T-Health-Condition corpus with those from the T-Health-Condition. From Table 1, the Drug-Health-Condition Baseline has the lowest recall of 63.86% of all the classification results and has the lowest precision of 25.67%. This indicates that the concurrent appearance of a drug name and a health condition in a tweet does not necessarily indicate an ADE. Model-7 outperformed all the models that used individual datasets (Model 2–5) and improved the F1-score by 2.53%. This shows that Models 2–5, which combine all the unlabeled data achieved superior results.

Tutubalina and Nikolenko focused on extracting phrases about ADRs from users' posts [43]. They applied a combination of RNN and CRF for sequence modeling by feeding word-level representations into a bidirectional RNN to encode the context vectors for each word and used a sequential CRF to jointly decode the words' labels for the entire sentence. Their joint model, as shown in Table 1, consistently outperformed the other models in terms of both precision and F-measure, while staying almost on par with the best RNN models in terms of recall. In conclusion, the combined GRU RNN + CNN + CRF model improves the quality of ADR extraction from free-text reviews. It also shows that concatenating input word embeddings with an extra embedding vector based on a character-level CNN significantly improves the results.

Current DL models generally utilize word embedding to learn similar representations for semantically similar words. In tweets or in biomedical text, a large number of words have no corresponding word embeddings; hence there is a challenge of dealing with out-of-vocabulary (OOV) words which cause misclassification. To deal with OOV words, Ding et al. introduced word- and character-level representations using a character embedding layer and word embedding layer for two Twitter datasets that had many informal vocabularies and irregular grammar and a biomedical text dataset extracted from PubMed abstracts with many professional terms and technical descriptions [9]. They then used an embedding-level attention mechanism to combine the two features to allow the model to dynamically determine

which information came from character- or word-level features. This embedding-level attention mechanism also enabled the model to decide how much information to use from character-level components and how much from word embeddings. A bidirectional gated recurrent unit (BiGRU) network was then used to encode the output derived from the attention layer. A separate output from the previous embedding-level attention layer, which was regarded as an auxiliary classifier, was concatenated with the output of the BiGRU layer to obtain the final output to improve the overall performance. They obtained competitive F1 scores of 0.844 and 0.906 for ADR identification on the Twitter and PubMed datasets, respectively [9], which is higher than the F1 score of 0.7549 achieved by Cocos et al. on the identification of ADRs from the Twitter dataset, as shown in Table 1.

Chu et al. developed context-aware attention mechanisms into the BiLSTM to detect adverse medical events (AMEs) from Chinese medical text that consisted of 8845 medical records with a total of 250,901 words of patients with cardiovascular diseases [6]. The model was different from other attention mechanism models as the model was able to calculate the attention signals by learning local context-aware information from a local part of the input data. The calculated attention signals were then assigned back to the original input data to give direct-viewing interpretability of the model. Compared with the baseline models, such as support vector machines (SVMs), BiLSTM and CNN, the experimental results demonstrate that this model achieves competitive performance in detecting AMEs from unstructured electronic health record (EHR) data because of its ability to learn local contextual information and capture the most pertinent information in the medical text [6].

Although previous attention approaches delivered promising results in the ADR classification task, they only extract single semantic information contained in single sentence representation [54]. To capture the different semantic information represented by different parts of a sentence, Zhang et al. developed a multihop self-attention mechanism (MSAM) model which employed multiple vectors for sentence representation to detect ADRs from two ADR datasets: TwiMed which contains sentences extracted from the Twitter corpus and the PubMed corpus as well as the ADE corpus extracted from 1644 PubMed abstracts [54]. Each attention step included in the model was designed to obtain different attention weights focusing on different segments to capture the multi-aspect semantic information for ADR detection. Their model achieved an F-measure of 0.853, 0.799 and 0.851 for ADR detection for TwiMed-PubMed, TwiMed-Twitter and ADE, respectively [54].

El-allaly et al. proposed a weighted online recurrent extreme learning machine (WOR-ELM) method to identify the boundary of ADR mentions from biomedical texts [10]. The proposed model consisted of two stages, namely, (a) span detection for identifying the boundary of the mentions irrespective of their type and (b) ADE mentions classification for classifying the identified mentions to the appropriate type. In both stages, they combined character-level and word-level embeddings as features for WOR-ELM. The character-level embedding was generated using a modified online recurrent extreme learning machine. It was then concatenated with word-level embedding to get the feature representation for each word in the vocabulary. The extracted features were then fed into the WOR-ELM to identify the mentions from a given input sentence and classify the identified mentions. Their method achieved an outstanding F-score of 87.5% [10].

From these aforementioned findings, we can conclude that both LSTM and GRU RNN are valuable tools for extracting ADRs.

Models with assigned context-aware attention weights, character and word embeddings, pre-trained with a suitable corpus give better results. The performance can be further enhanced if word embedding is done across different types of corpora. A joint model which is a combination of CNN, LSTM, RNN and CRF with assigned attention weights, fed with word embedding datasets, may deliver outstanding results and can be an excellent model for ADR extraction.

Extraction of drug–drug interactions

To address the ADR problem, in addition to investigating ADRs caused by the consumption of a single drug, it is appropriate to examine many other factors that bring about ADRs. ADRs can be predicted by firstly (a) examining the structural similarities of compounds and proteins and secondly (b) by examining DDIs. While the former is mainly used to predict ADRs caused by new drugs based on ADRs of existing drugs, the latter requires comprehensive studies on drug–target interaction. Since DDI is one of the causes that can lead to unexpected ADRs, the detection and prediction of ADRs induced by DDIs are very important [49].

Many methods have been proposed for DDI extraction. Both ruled-based and machine learning methods have been used [39]. Generally, the application of deep learning approaches in DDI extraction can be grouped into two types, namely, binary classification and multi-class classification [60]. The former is to determine whether a DDI exists between two drugs, while the latter is to determine the type of DDI, namely, ADVICE, EFFECT, INT and MECHANISM [13]. The dataset is annotated with the following type of interactions:

Advice: The sentence is an opinion or recommendation-based sentence.

Effect: The sentence indicates the effect of drug–drug interaction which describes the pharmacodynamics mechanism of drug interaction.

Mechanism: The sentence describes the pharmacokinetic mechanism.

Int: The sentence only indicates a drug interaction without giving any other detailed information.

Details of different DDI types can be found in Table 2.0.

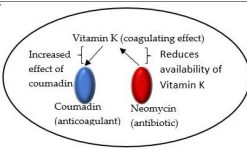
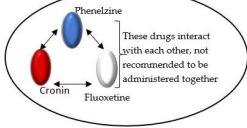
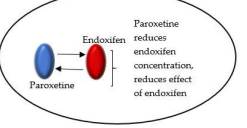
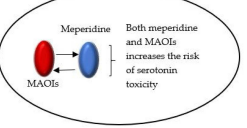
Deep learning models used for DDI classification

Seeing that CNN is a robust machine learning method which does not need manually defined features, Liu et al. used CNN for DDI extraction on the DDI corpus of the 2013 DDIExtraction challenge [27]. This was the first time that CNN was used for DDI extraction. Their CNN-based system outperformed all other systems, most of which were based on SVM with various features such as syntactic features and features derived from external lexical resources. Table 3 summarizes the deep learning methods used to detect DDIs from specialized databases and the scientific literature.

To solve the messy feature engineering problems, Quan et al. [36] utilized multichannel CNN to address the ‘vocabulary gap’ problem, and Zhao et al. used syntax convolutional neural networks (SCNNs) for DDI extraction (Table 3) [58]. Their experiment results on the DDIExtraction 2013 corpus showed that SCNN achieved better performance (F-score of 0.686) than other traditional machine learning methods.

Riding on the attention-based concept, the LSTM model consisting of a general pooling with attention was then used for DDI

Table 2. Illustration of different drug–drug interaction types

Categories	Sentence	Annotation	Sample
Effect	Oral neomycin sulfate (drug 1) may enhance the effect of coumadin (drug 2) by decreasing vitamin K availability (Zhou et al., 2018) [60]	Drug 1 and drug 2 in the above sentence falls into EFFECT category because the sentence shows pharmacodynamics effect of interaction.	
Advice	Fluoxetine (drug 1) and crocin (drug 2) should not be administered to patients receiving phenelzine (drug 3) (Sahu and Anand, 2018) [37]	Both interactions between drug 1 & drug 3 and drug 2 & drug 3 fall under ADVICE category. Advice has been given not to take drug 1 and drug 3 together and drug 2 and drug 3 together.	
Mechanism	Paroxetine (drug 1) reduces the plasma concentration of endoxifen (drug 2) by about 20% (Sahu and Anand, 2018) [37]	The interaction between drug 1 and drug 2 falls under MECHANISM category because the sentence describes a pharmacokinetic mechanism.	
Interaction	This is the typical interaction of meperidine (drug 1) and MAOIs (drug 2).	No other information is provided in the above sentence except touching on drug interaction only.	

classification [59]. However, the introduction of pooling attention failed to improve the performance of DDI classification. The model for DDI classification was then improved using an attention-based neural network model (Att-BLSTM) that combined an attention mechanism and a recurrent neural network with LSTM units [51]. The word with its part-of-speech (POS) tag obtained from the Stanford Parser [8] was used to distinguish the semantic meaning of different sentences [51]. Basically, this model consisted of (a) an input layer which carried the words, POS and relative distances between a word and each candidate drug in an input sentence; (b) the embedding layer to encode the input into real-valued vectors known as embedding vectors; (c) the input attention layer to weight word embedding vectors that could identify the relationship between a pair of special candidate drugs; (d) the concatenating layer to connect three embedding vectors into a vector by words; (e) a bidirectional RNN with an LSTM unit to learn the high-level syntactic meaning of the whole sentence and pass the outputs over to the last layer; and (f) a logistic regression layer with the softmax function to perform DDI classification. This model achieved a high detection rate and classification score, which were higher than other state-of-the-art methods. It is evident from Table 3, Figures 2 and 3 and that generally deep learning neural networks that are equipped with the LSTM model and enhanced with special features such as Att-BiLSTM model (Table 3) outperform CNN models in precision, recall and F1-score. This is due to the fact that the LSTM network outperforms other models in processing long sequential data. A further comparison between the LSTM-based models and CNN-based models reveals that both LSTM-based and CNN-based models give comparatively good results in relation to the precision value for all types of detections and LSTM-based models give much better results in recall and F1-score values (Table 3). Figure 3A–D shows that CNN-based models perform badly in relation to recall values for all types of extraction, especially on DDI Int extraction (Figure 3D). This could be due to the fact that insufficient training data was used and hence contributed to imbalanced outcomes.

Table 3. Summary of the deep learning methods used for DDI extraction using DDI corpus of the 2013 DDIExtraction challenge

Study	Techniques	Precision	Recall	F1-score
Zhou D et al. [60]	PM-LSTM: Used position-aware in LSTM network for deep multitask learning	75.80	70.38	72.99
	P-BiLSTM: Used position-aware in LSTM network for DDI classification	74.57	70.07	72.25
	M-BiLSTM: Developed multitask learning framework for DDI multi-classification	71.90	71.60	71.75
Zhang Y et al. [57]	BiLSTM and 2CNN: Hybrid model which combined BiLSTM RNNs and 2CNNs for the extraction of DDI	77.1	73.3	75.1
Zheng W et al. [53]	Att-BiLSTM: Incorporated an attention mechanism into BiLSTM to automatically learn more influential words	78.4	76.2	77.3
Zhang Y et al. [59]	HRNNSDP: Used hierarchical recurrent neural networks (HRNN) to integrate the shortest dependency path (SDP) between the two drugs for DDI extraction	74.1	71.8	72.9
Asada M. et al. [1]	AttCNN: A separate attention mechanism was incorporated into the model and a bias term was incorporated to adjust the smoothness of attentions	76.30	63.25	69.12
Sahu and Anand [37]	BiLSTM: Utilized word and position embedding. Bidirectional long short-term memory (BiLSTM) networks were used in DDI extraction.	75.97	65.57	70.39
	ABiLSTM: Used word and position embedding; used attentive pooling on the outputs of BiLSTM to get fixed length features over complete sentence	67.85	65.98	66.90
	Joint ABiLSTM (BiLSTM+ ABiLSTM): Max pooling was applied on the first BiLSTM and attentive pooling was applied on the second BiLSTM layer	73.41	69.66	71.48
Yi Zibo et al. [51]	BiLSTM + 2ATT: This model had multiple attention layer model. It consisted of word-level attention to transform sentence matrix to vector representation and sentence-level attention layer to generate final representation	73.67	70.79	72.20
Wang W. et al. [47]	DBiLSTM: Three channels, namely, Linear channel, Depth First Search channel and Breath First Search channel, were constructed with three network layers (embedding layer, LSTM layer and max pooling layer) from bottom up. The embedding layer extracted distance-based feature and dependency-based feature	72.53	71.49	72.00
Zhao Z et al. [58]	SCNN (The syntax convolutional neural network): Used syntax word embedding and dependency tree to capture syntactic information of a sentence	72.50	65.10	68.60
Quan et al. [36]	MCCNN: The multichannel convolutional neural network (MCCNN) model utilized word embedding, especially the multichannel word embedding layers to address 'vocabulary gap' problem and integrate semantic information	76	65.3	70.2
Liu S et al. [26]	CNN + DCNN: To take advantage of long-distance dependencies between words, dependency-based convolutional neural network (DCNN) was used for DDT extraction. CNN and DCNN were combined to reduce error propagation	78.24	64.66	70.81
Liu S et al [27]	CNN: Deployed CNN-based method for DDI extraction	75.72	64.66	69.75

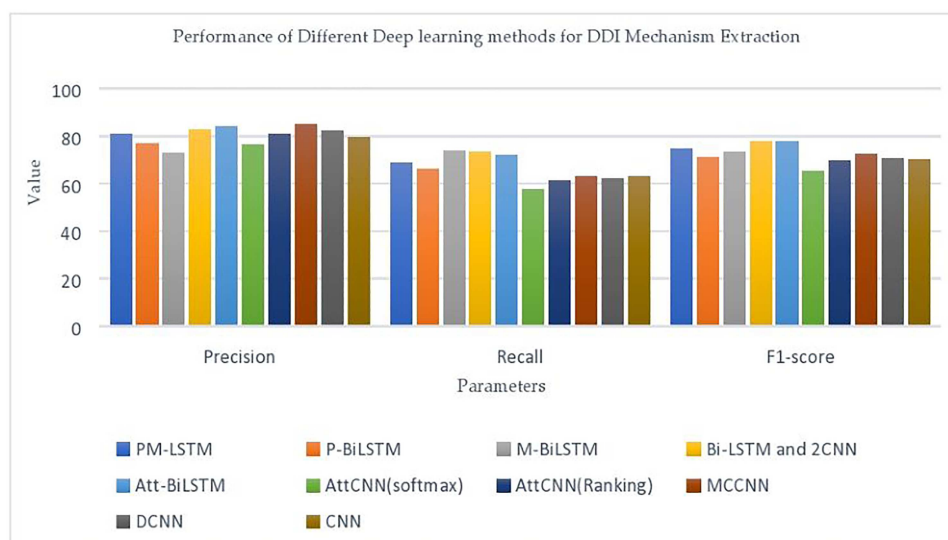


Figure 2. Comparison of the performance of different deep learning models used in DDI extraction. Three parameters (precision, recall and F1-score) were used as comparison tools for the measurement of performance of different deep learning methods for DDI mechanism extraction. The results show that Att-BiLSTM was the most effective in precision, recall and F1-score, followed by BiLSTM and 2CNN models.

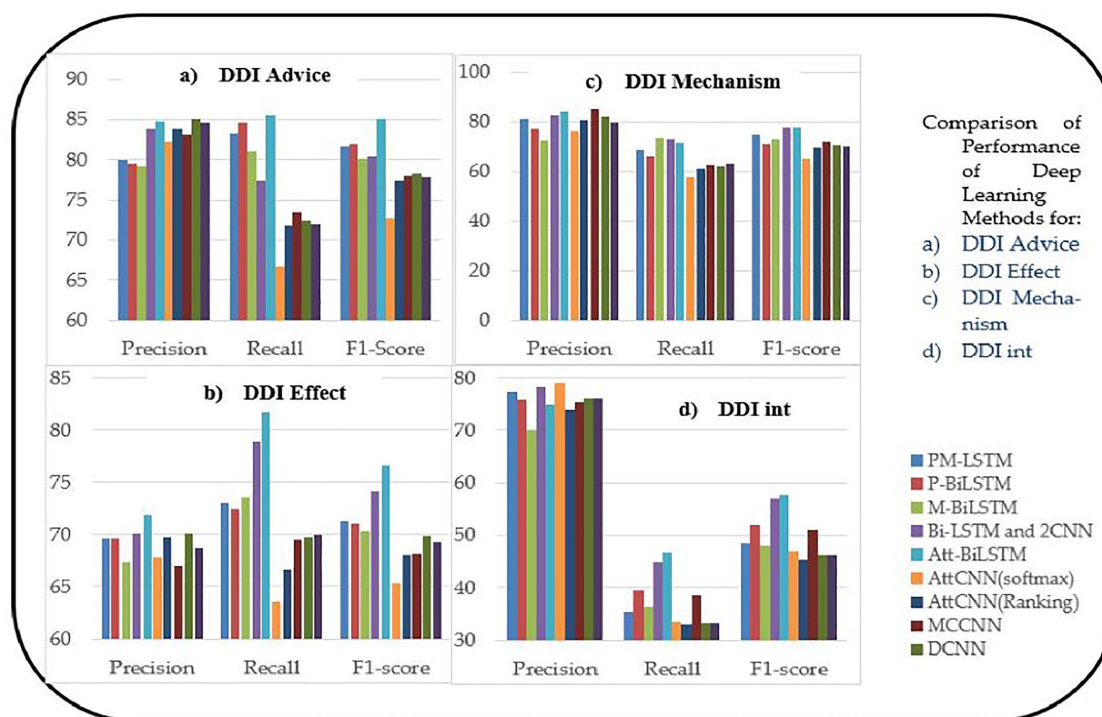


Figure 3. Comparison of the performance of different deep learning models in four DDI (drug-drug interaction) classifications, namely, DDI Advice, DDI Effect, DDI Mechanism and DDI int. The parameters used for the comparison were precision, recall and F1-Score. The results in Figure 3A–D show that Att-BiLSTM is the most effective model with high values in precision, recall and F1-score.

From Tables 3 and 4, it is evident that all the past research studies focus solely on using different deep learning models for DDI detection and classification. The results show that the performance of generic CNN and RNN variant models can be strengthened by enhancing the model with extra features such as word embedding, position embedding, the attention mechanism, dependency mechanism and other features. They also show that some models excel at balancing the precision and

recall values while some others achieve lower success in balancing the values.

Deep learning models for DDI prediction

Most of the aforementioned approaches characterize DDIs using scores to represent the overall strength of an interaction and

Table 4. Hyperparameters used in deep learning methods for DDI detections

Study	Zhou et al. [54]	Zhang et al. [51]	Zheng et al. [53]	Zhang et al. [50]	Asada et al. [1]	Sahu and Anand [32]	Yi et al. [45]	Wang et al. [40]	Zhao et al. [52]
Input layer: Embedding layer	Word and position embeddings	Word and position embeddings	Word, position and POS embeddings	POS and position embeddings	Word embedding	Word and position embeddings	Word and position embeddings	Syntax word and random word embeddings	Syntax word, POS and position embeddings
Embedding software	Pre-trained unlabeled biomedical texts crawled from PubMed using Word2vec	Used Word2vec for word embedding. Used dependency relation and position embedding	Word2vec	Word2vec for word embedding and Stanford parser to get the dependency syntax relations	Skip-gram was employed for the pre-training of word embeddings	GloVe was used on a corpus of PubMed open source articles	Deployed GloVe	Word2Vec	Enju parser and Word2Vec
Deep learning model	Position attention-based BiLSTM multitask learning	Combination of BiLSTM and 2CNN	Incorporated an attention mechanism into BiLSTM	Hierarchical recurrent neural networks (HRNN)-based method	CNN attention-based DDI extraction model (AttCNN) with a separate attention mechanism incorporated into the model	Three LSTM-based models: BiLSTM used max pooling; AB-LSTM used attentive pooling; Joint AB-LSTM	BiLSTM gated recurrent unit (GRU) with multiple attention layers instead of max pooling	Deployed dependency-based bidirectional LSTM model with three channels (Linear, DFS and BFS channels)	
Hidden unit dimension	150	300	230	100	100	150 and 200	220	300	CNN was used
Dropout to avoid overfitting	Used dropout in the embedding, BiLSTM and the output layers	Dropout rate after embedding and output layers is 0.3	Dropout rate of 0.5 was applied	The dropout rate of embedding layer and output layer were set as 0.7 and 0.5	NA as CNN was used. Initial learning rate was 0.001	Applied dropout on the output of the pooling layers at 0.7 and 1.0	The probability of dropout was 0.5	0.7	CNN was used
Output layer	Two softmax classifiers: for binary and for multi-class classification	A max pooling was applied. MLP model combined the outputs of RNNs and CNNs	Used softmax layer to perform DDI classification	Softmax function was used as the activation function of the output layer for classification of DDI	Used two different objective functions, softmax and ranking for classification	Used softmax function in the output of fully connected neural layer to normalize probability score	Used softmax classifier. Used cross-entropy function and L^2 regularization as the optimization objective	The output was concatenated together and fed to the softmax layer for classification	Fed to softmax classifier to extract DDIs

do not provide specific descriptions of DDI in terms of pharmacological effects. In view of the fact that the aforementioned methods failed to provide sufficient details beyond the chance of DDI occurrence, Jae et al. developed a new computational framework, known as DeepDDI, to study drug–drug interactions [16]. By using drug–drug constituent pairs and with the drugs' structural information as inputs, they accurately generated 86 DDI types with a mean accuracy of 92.4% using the DrugBank gold standard DDI dataset covering 192,284 DDIs contributed by 191,878 drug pairs. They also used DeepDDI to discover potential causal mechanisms for 9284 drug pairs. Their DeepDDI was able to provide a better understanding of the DDI mechanism which caused drug interactions, and it was able to suggest alternative or replacement drug candidates for 62,707 drug pairs which had negative health effects [16].

DeepDDI predicted multiple DDI types for given drug pairs by simply using drug names and the chemical structures of drug pairs in SMILES as inputs fed into DNN. The model employed an optimized DNN along with two structural similarity profiles (SSP) generated as the feature vectors of each drug pair. After being subjected to dimension reduction, the two SSPs which were combined as a single SSP vector of a drug pair gave outputs which showed high accuracy, suggesting that DeepDDI provided more specific information on drug interactions in sentences which were human readable and interpretable, beyond simply predicting the occurrence chance of DDIs only. However, the identified side effects are predominantly linked to SSP and cannot be associated and explained in terms of a molecular basis. Moreover, the SSP can be correlated with side effects only to a certain extent as many drug pairs have high structural similarity but do not have similar side effects [31].

In the same year, Zitnik et al. used a very different approach to predict the side effects of a drug–drug combination. They cast the problem of predicting polypharmacy side effects as similar to solving a multi-relational link prediction task [61]. They built and encoded the multi-relational link among drug, protein and polypharmacy side effects on a two-layer multimodal graph.

Their approach, Decagon, showed improvement in DDI prediction as the protein–protein interaction (PPI) network was able to consider physical interactions such as metabolic enzyme-coupled interactions and signaling interactions that were experimentally documented in humans. Moreover, they also included relationships between proteins and drugs from the STITCH database and the side effects of both individual drugs and drug combinations from SIDER, hence successfully integrating various chemical and protein networks. Their graph convolutional model achieved excellent accuracy on the polypharmacy side effect prediction task on nearly a thousand different side effect types by integrating molecular and patient population data, providing insights into the clinical manifestation of DDIs [61].

Challenges faced in using deep learning approaches

DL methods are complex in nature because the networks contain numerous hidden layers and millions of neurons interacting with each other. Despite its great advancement in biomedical applications for the detection, classification and prediction of ADRs and DDI, there are still numerous challenges facing these DL approaches.

a. Non-interpretable biological prediction

From the review, it is clear that ADRs/DDI classification approaches can be categorized into classification-based and similarity-based methods. Many researchers have applied deep learning approaches to perform both the binary and multi-classification of DDI. All these published deep learning methods successfully detected ADRs and the likelihood of DDIs by identifying the occurrence of DDI and classifying them into four categories. They only give a scalar explanation and do not give an interpretable biological prediction at a molecular level, which is one of the biggest challenges faced in DL methods.

Similarity-based methods such as using structural similarity profile measures for DDI prediction were successful to a certain extent [16]. However, prediction models based on the structural similarity of drug pairs are not sufficient to give an accurate prediction because each drug may exert some effects on human bodies, and the two drugs may be metabolized separately via two remarkably distinct pathways, producing two groups of metabolites that differ greatly in chemical structure, targets, and physiological effect [50, 55]. Moving forward, the science of implementing deep learning must be improved, and interpretable deep learning algorithms must be developed to unfold the black box characteristic of DL.

b. Expert-dependent algorithm development

The availability of a big medical data and advancements in computing power have contributed to the popularity of DL in the biomedical field. These two factors have enabled the DL to learn hierarchical features and representations of big data with less feature engineering work. However, a big dataset poses a challenge when training the DL model because in order to learn numerous features of a large amount of data, experts must develop advanced algorithms for parallel DL models and GPU-based implementation to optimize the performance of deep learning models [52]. With the continuing growth in big data, new learning frameworks which can compress large-scale DL models and advanced computing infrastructure must be further developed to improve training efficiency and prediction accuracy.

Diverse data provide different and useful information for ADR and DDI predictions. Heterogeneous data sources may contain image and text simultaneously. Each modality of multi-model objects has different characteristics and hence adds complexity to training the DL model. Recently, a deep computation model was developed to process heterogeneous data, and the model achieved 2–4% higher classification accuracy than multimodal DL models [52]. Again, to explore more effective fusion methods to improve the multimodal DL models, a lot of expert effort is required to advance this area.

From the review, it is obvious that low-quality data such as incomplete sentences, noise, imprecise expressions and redundant objects are present in big data, especially in social media data sources. With the explosion in the growth of big data, it is foreseeable that more research effort is needed to develop reliable DL models for processing huge, heterogeneous and low-quality data to achieve the better detection and prediction of ADRs.

Future direction for predictions of ADRs

There has been a significant paradigm shift from using traditional statistical and machine learning models to deep learning models for the detection and prediction of ADRs. Current deep learning models have established a solid foundation for the

detection and classification of ADRs induced by individual drugs and DDI. When compared to the baseline (CFR-context) scores shown in Table 1, all the scores of the RNN models showed improvement over the baseline scores. The GRU-document had a recall of 0.8126, precision of 0.7938 and F-score of 0.8031, which was an improvement of 19, 2 and 11%, respectively over the recall, precision and F-score of the baseline [17]. A similar trend can be seen in the scores achieved by the four-layer GRU which had a recall of 0.7262, precision of 0.6565 and F-score of 0.6896, which was an improvement over recall of 0.5972, precision of 0.6254 and F-score of 0.6110 of the baseline CRF [43]. From the score, we can conclude that the DL models outperform the baseline models and have established a solid foundation for the detection and prediction of ADRs.

In view of the importance of using deep learning approaches to help reduce and prevent the occurrence of ADRs, we propose several emerging factors that can help to enhance the accuracy and robustness of deep learning for the future prediction of ADRs.

(a) Integrating big data

Generally, to the best of our knowledge, all the deep learning prediction models adopted for DDI prediction so far are applicable only to marketed drugs with available side effect information, detailed PPIs, DDIs and drug-protein interaction network information. To facilitate the prediction of ADRs induced by multiple drug interaction (more than a two-drug interaction) or to predict rarer side effects induced by DDI, there is a need to tap into big data to discover more unknown facts.

A vast amount of biomedical resources and heterogeneous healthcare data which carry multidimensional drug properties offer new opportunities for detecting and predicting ADRs caused by both single drugs, DDIs and multiple drug interaction. The efficacy of deep learning in acquiring different ADR information across different domains on a cross-domain platform is a promising but largely unexplored area.

Recently, a very new architecture, known as a differentiable neural computer (DNC), has been used in deep learning memory-augmented neural networks [5]. The structure of DNCs is more robust and abstract, and it can perform tasks that have longer-term dependencies than some of its predecessors such as the LSTM network [14]. It combines the learning and pattern-recognition strengths of deep neural networks with the ability to retain information in complex data structures such as graphs in a computer memory. DNCs have been applied to question-answering systems and find the shortest path in graphs. We believe that DNCs can be trained to perform complex, structured tasks and address big data applications such as performing better semantic text analysis to give better predictions of ADRs caused by both single drugs and drug-drug combinations.

(b) Developing innovative deep learning models for the detection and classification of ADRs and DDI

Zhang et al. provided a comprehensive review of deep learning concepts and models [53]. It was reported that deep learning techniques can be a single deep learning technique, a deep composite model or an integrated model. Different deep learning models have different strengths [23, 46, 53]. For example, CNN is capable of extracting local and global representations from heterogeneous data sources such as textual and visual information. On the other hand, RNN can process the sequential influences of content information. The deep semantic similarity model is able to perform semantic matching between words. These models

complement one another and become a more powerful deep composite model.

The detection of ADRs is a sequence labeling task, and most studies use either CNN or BiLSTM to perform ADR detection (Tables 3 and 4). Additional mechanisms such as the attention mechanism are applied to CNN or BiLSTM to enable the model to process long and noisy inputs [5], as shown in Figure 4. Although BiLSTM is effective in capturing useful information within a sentence and producing a representation of the entire sentence after applying a max pooling layer, when one sentence contains two or three events or more, existing models have problems in capturing the interaction between candidate arguments as all these candidate arguments must be linked to capture the valuable semantics of the whole sentence. Traditional models such as traditional BiLSTM and CNN approaches do not model the interactions between candidate arguments and predict them jointly, so it is worth considering strengthening the existing models further by developing many more innovative learning models in the near future.

Initiatives such as exploring the possibility of integrating the neural networks of BiLSTM and CNN, which allows BiLSTM to extract long short-term dependent information and utilizes a two-dimensional convolution to extract and represent the syntactic structural information of the text, and finally using one-dimensional pooling to process the feature vector of the matrix are worth considering. The addition of appropriate useful mechanisms into suitable existing deep learning models (Table 3), such as adding an attention-based mechanism [4, 59] or other more innovative mechanisms, enhanced with optimized hyperparameters (Table 4) to capture the informative elements of the inputs, can assist in providing better interpretability and prediction.

(c) Facilitating biological interpretability

Biological interpretability is very important because the response of the human body to drugs is a complex phenomenon. Every drug interacts differently with the human body. The effectiveness or ineffectiveness of a drug on different persons with different inherent biological properties varies. Very often, healthcare providers fail to foresee an ADE that may happen to patients who have had no previous ADR experiences. To prevent healthcare providers from merely administering drug prescriptions based on what has worked well for other patients in the past to reduce the occurrence of ADE, the mechanisms and reasons why ADRs occur must be clearly tabulated.

The successes of CNN in image recognition [18] and RNN LSTM in natural language processing, speech recognition and machine translation [52] can be leveraged to find chemical substructures which are associated with ADRs. The feature learning capability of CNN and LSTM indicates that the networks can potentially learn the relationship between a drug's molecular structure and a drug's property accurately. Neural networks can infer the topology of drugs directly from the simplified molecular input line entry system (SMILES) branching and ring closures defined by the brackets and numbering in the SMILES strings. Leveraging the potential capacity of DL to perform complex visual spatial reasoning tasks such as decomposing or breaking a drug molecular structure into a series of smaller drug molecular substructures, those substructures which cause the side effects can then be potentially deduced to predict side effects, resolving the black box issue or 'non-interpretability' issue of DL methods, and finally predictions can be made in readable sentences. This

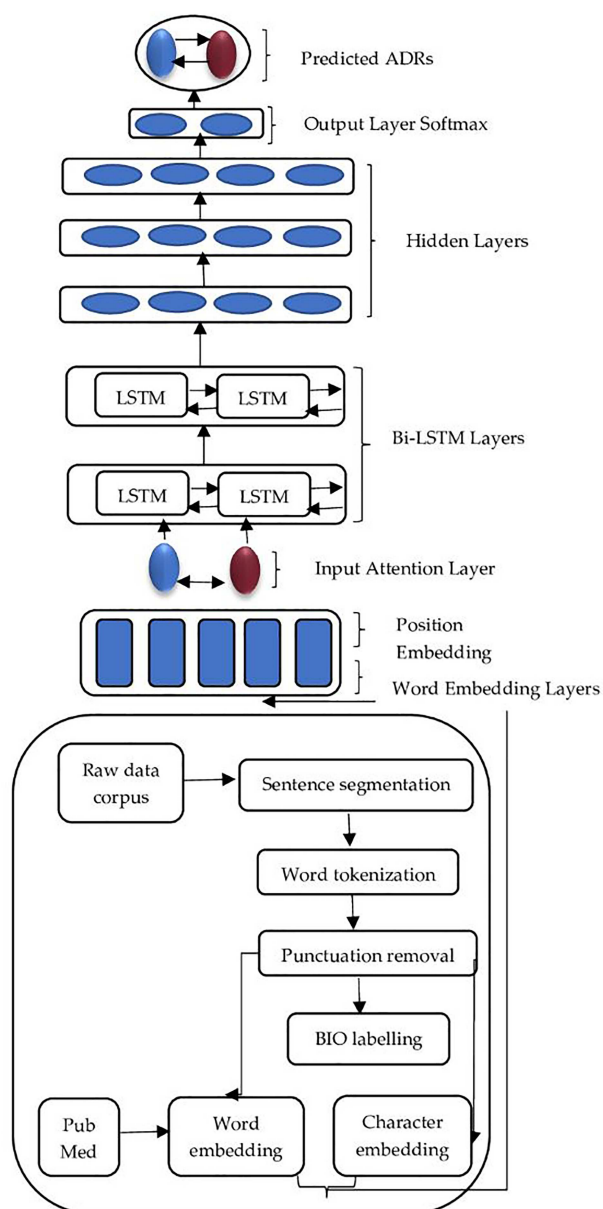


Figure 4. The architectural framework of the BiLSTM model commonly used for detecting ADRs.

will help to prevent ADRs in the early stage of the drug development pipeline which in turn will help to enhance drug safety and reduce costs.

From ADRs prediction to drug repurposing

It is general knowledge that the drug discovery process is time-consuming, and an enormous amount of money and effort is needed to develop a new drug. Drug repurposing means repurposing a tested drug which has been approved for other uses. This has been used as a strategy to shorten the drug discovery process. Once the drug is repurposed, the new drug candidate can be sent for clinical trials quickly and can be reviewed by the Food and Drug Administration for approval.

One of the methods that has been used to repurpose drugs is based on drug similarity. Drug chemical structural similarity

and phenotypic similarity at a gene expression level or at the whole organism level have been applied widely to infer drug targets. The underlying assumption is that if two drugs are in some way similar enough, then they are likely to share targets and physiological effects, thus having the potential to treat the same diseases. As DL methods are capable of discovering latent and complex structures in large datasets, they can be used to integrate drug chemical data, protein interaction, pathway and biological ontological data as well as other multiple sources of data. By using backpropagation algorithms and adjusting connecting weights, DL methods can compute the representation of each layer based on that of the previous layer for ADR prediction. If designed correctly, in addition to predicting ADRs, the effective integration of many different types of data can also infer a link between a drug and a new protein target, making connections to identify a new target that is not likely to be seen by researchers on their own, highlighting a large unexplored niche for drug-target interaction prediction in drug repurposing.

Traditional approaches usually focus on only one aspect of the problem such as common side effects or adverse drug reactions and fail to consider the complexity of diseases and their mechanisms of action or the complexity of patient populations. DL repurposing methods can be extended to finding new indications for individual patients when patients' heterogeneity and complexity are included as input components, hence reducing the risk of drug toxicity or inefficacy caused by interpatient variability. DL methods that use 1D representations of targets and drugs have been shown to be an effective approach for drug-target binding affinity prediction [34]. As DDI data may contain information on drug targets or physiological effects, the data can provide some new ideas for drug repurposing. Till now, DDI has been somewhat unexploited for drug repurposing. In our opinion, DDI data are complementary to other drug information, and drug repurposing based on drug-drug interaction data using deep learning approaches is feasible. Drugs which have safe pharmacokinetic, pharmacodynamics and toxicity profiles and have been commercially available in the market can then be readily sent to Phase II and Phase III clinical studies, and this will result in a decrease in development cost, a better return on investment and an accelerated development.

From prescription advice to nutrition advice

In addition to ADRs caused by single drugs and DDIs, ADRs can also be caused by drug-supplement interactions and drug-food interaction. Some dietary supplements and food may interact unfavorably with drugs, and this may magnify the effect of ADRs. For example, some drugs can deplete vital nutrients that are critical for good health and thereby may induce ADRs and diseases. On the other hand, dietary supplements such as calcium can lower the absorption of antibiotic tetracycline, hence reducing the drug's therapeutic potential. Similarly, herbal supplements which contain active potent ingredients may interact with drugs or with food or alcohol. Deep learning approaches will open the door for a better understanding of ADRs induced by drug-drug-food, drug-drug-supplements through the integration of dietary supplement information and daily food consumption type information. This will allow healthcare practitioners to offer suggestions to use alternative drug candidates for better disease treatment. Healthcare practitioners can also provide dietary advice regarding food nutrition and vitamin supplements to patients, especially those who are involved in polypharmacy to prevent the occurrence of ADRs, hence reducing the healthcare costs incurred by ADRs.

Conclusion

In this paper, we provide a review on the application of deep learning in the detection of ADRs caused by single drugs and DDIs. We provide a detailed presentation on the deep learning approaches used in each case and give a clear understanding of the deep learning concepts together with their advancement in extracting ADRs. Such an extensive review is important for future research, as researchers and practitioners will have a better understanding of the strengths and weaknesses as well as the application scenarios of each model, enabling them to quickly step into the field of applying deep learning methods in detecting and predicting the potential occurrence of ADRs.

We also highlight the importance of utilizing heterogeneous data sources and innovative deep learning methods to detect ADRs. We identify future directions and propose to use more robust, state-of-the-art deep neural network methods to predict ADRs induced by single drugs or drug–drug interaction, to carry out drug repurposing and predict drug–drug–food interaction by integrating chemical, biological, phenotypic, network and many other types of data into the prediction process. We are of the opinion that the development of deep learning-based models using multidimensional drug properties and heterogeneous data sources is a promising strategy to predict unknown ADRs, especially those induced by multiple drug–drug combinations. The deep learning approaches are tools which cannot be ignored, and their applications must be explored and extended to other important healthcare domains, such as drug repurposing and drug–drug–food interaction.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa040/5826453>

Key Points

- Application of deep learning approaches in extracting ADRs caused by single drugs and drug–drug combinations
- Integration of multidimensional drug properties across heterogeneous data sources to detect and predict ADRs induced by DDIs using deep learning approaches
- Deployment of deep learning approaches for replacements for drugs which have side effects with drugs with reduced side effects
- Diversification the utilization of drugs through drug repurposing
- Challenges faced in using deep learning for prediction of side effects
- Emerging factors that can help to enhance the accuracy and robustness of deep learning models for ADR prediction

References

- Asada M, Miwa M and Sasaki Y. Extracting drug–drug interactions with attention CNNs. In *Proceedings of the BioNLP Workshop*, pp. 9–18, Canada.
- Atias N, Sharan R. An algorithms framework for predicting side effects of drugs. *J Computing Biol* 2011;18:207–18.
- Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on neural network* 1994;5:157–66.
- Cao Z, Wang L, Melo DG. Multiple-weight recurrent neural networks. In *Proceeding of the Twenty-Sixth International Joint Conference on Artificial Intelligence* 2017, pp. 1483–9.
- Chen HM, Engkvist O, Wang YH, et al. The rise of deep learning in drug discovery. *Drug Discov Today* 2018;23(6):1241–50.
- Chu J, Dong W, He K, et al. Using neural attention networks to detect adverse medical events from electronic health records. *J Biomed Inform* 2018;87:118–30.
- Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association* 2017;24:813–21.
- De Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: *Proceedings of LREC*, 2006; pp. 449–54.
- Ding P, Zhou XB, Zhang XJ, et al. An attentive neural sequence labelling model for adverse drug reactions mentions extraction. *IEEEAccess* 2018;6:73305–15.
- El-allaly E, Sarrouiti M, En-Nahnah N, et al. An adverse drug effect mentions extraction based on weighted online recurrent extreme learning machine. *Comput Methods Programs Biomed* 2019;176:33–41.
- Goodfellow I, Bendigo Y, and Courville A. *Deep Learning*. MIT Press, Cambridge, MA; 2016. [Online]. <http://www.deeplearningbook.org>.
- Gupta S, Pawar S, Ramrakhiani N, et al. Semi-supervised recurrent neural network for adverse drug reaction mention extraction. *BMC Bioinformatics* 2018;19(Suppl 8):212.
- Herrero-Zazo M, Segura-Bedmar I, Martinez P, et al. The ddi corpus: an annotated corpus with pharmacological substances and drug–drug interaction. *J Biomed Inform* 2013;46:914.
- Hsin C. Implementation and optimization of differentiable neural computers. *Technical report*. Stanford University. 2016.
- Huynh T, He Y, Willis A et al. Adverse drug reaction classification with deep neural networks. In: *The 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 877–87, Osaka, Japan.
- Jae YR, Hyun UK, Sang YL. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc Natl Acad Sci* 2018;115(18):4304–11.
- Jagannatha AN and Hong Y. Bidirectional RNN for medical event detection in electronic health records. *Proceedings of NAACL-HLT*, 2016, pp. 473–82.
- Jing YK, Bian YM, Hu ZH, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The American Association of Pharmaceutical Scientists* 2018;20:58.
- Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;38:D355–60.
- Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;44:D1075–107.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- Lee K, Qadir A, Hasan SA et al. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. *Proceedings of the 26th International Conference on World Wide Web*, vol. 2017, pp 705–14.

23. Li F, Zhang Y, Zhang M et al. Joint model for extracting adverse drug events from biomedical text. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, 2016, pp. 2838–44. AAAI Press, Palo Alto.
24. Lipton ZC, Kale DC, Elkan C et al. Learning to diagnose with LSTM recurrent neural networks. In: *International Conference on Learning Representations*, 2015, pp. 1–18. San Diego, CA, USA.
25. Liu R, Mohamed DWA, Kumar K, et al. Data-driven prediction of adverse drug reactions induced by drug-drug interactions. *BMC Pharmacol Toxicol* 2017;**18**:44.
26. Liu S, Chen K, Chen Q, et al. Dependency-based convolutional neural network for drug-drug interaction extraction. In: *IEEE International Conference on Bioinformatics and Biomedicine*, 2016, pp. 1074–80.
27. Liu S, Tang B, Chen Q, et al. Drug-drug interaction extraction via convolutional neural networks. *Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine* 2016;**2016**:6918381.
28. Luo Y. Recurrent neural networks for classifying relations in clinical notes. *J Biomed Inform* 2017;**72**:85–95.
29. Magro L, Moretti U, Leone R. Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. *Expert Opin Drug Saf* 2012;**11**:83–94.
30. Mikolov T, Chen K, Corrado G et al. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*, 2013. Scottsdale.
31. Miotto R, Wang F, Wang S, et al. *Deep Learning for Healthcare: Review, Opportunities and Challenges*. Oxford University Press Briefings In Bioinformatics, 2017, 1–11.
32. Moore TJ, Cohen MR, Furberg CD. Serious adverse drug events reported to the Food and Drug Administration, 1998–2005. *Arch Intern Med* 2007;**167**:pp1752–9.
33. Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med* 2019;**14**(10):1–11.
34. Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics* 2018;**34**(2018):i821–9. doi: [10.1093/bioinformatics/bty593](https://doi.org/10.1093/bioinformatics/bty593).
35. Pierre LR, Lexchin J, Simonyan D. Analysis of the drugs withdrawn from the US market from 1976 to 2010 for safety reasons. *Pharmaceutical Medicine* 2016;**30**(5):277–89.
36. Quan C, Huo L, Sun X, et al. Multichannel convolutional neural network for biological relation extraction. *Biomed Res Int* 2016;**2016**:1850404.
37. Sahu SK, Anand A. Drug-drug interaction extraction from biomedical text using long short term memory network. *J Biomed Inform* 2018;**86**(2018):15–24.
38. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions* 1997;**45**(11):2673–81.
39. Segura-Bedmar I, Martinez P, Pablo-Sanchez CD. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC bioinformatics* 2011;**12**(Suppl 2):S1.
40. Szklarczyk D, Santos A, Mering CV, et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**:D380–4.
41. Tan Y, Hu Y, Liu X, et al. Improving drug safety: from adverse drug reaction knowledge discovery to clinical implementation. *Methods* 2016;**110**(2016):14–25.
42. Tatonetti NP, Ye PP, Daneshjou R, et al. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**:125ra31.
43. Tutubalina E, Nikolenko S. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *J Healthc Eng* 2017;**2017**:9451342.
44. Urban G, Bache KM, Phan D, et al. Deep learning for drug discovery and cancer research: automated analysis of vascularization images. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(3):1029–1035.
45. Vilar S, Harpaz R, Uriarte E, et al. Drug-drug interaction through molecular structure similarity analysis. *J Am Med Inform Assoc* 2012;**19**(6):1066.
46. Wang L, Cao Z, de Melo G et al. Relation classification via multi-level attention CNNs. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics*, 2016; pp. 1298–1307.
47. Wang W, Yang X, Yang C, et al. Dependency-based long short term memory network for drug-drug interaction extraction. *BMC Bioinformatics* 2017;**18**(Suppl 16):578.
48. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.
49. Yang H and Yang C. Discovering Drug-Drug Interactions and Associated Adverse Drug Reactions with Triad Prediction in Heterogeneous Networks. *IEEE International Conference on Healthcare Informatics*, 2013, pp. 244–54.
50. Yamanishi Y, Kotera M, Kanehisa M, et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**(ISMB 2010):246–54.
51. Yi Zibo, Li S, Yu J et al. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. *International Conference on Advanced Data Mining and Applications*, 2007, pp. 554–66.
52. Zhang QC, Yang TL, Chen Z, et al. A survey on deep learning for big data. *Information Fusion* 2018;**42**(2018):146–57.
53. Zhang S, Yao L, Sun A. Deep learning based recommender system: a survey and new perspectives. *ACM Comput Surv* 2019;**1**(1):1–38.
54. Zhang T, Lin HF, Ren YQ, et al. Adverse drug reaction detection via a multihop self-attention mechanism. *BMC Bioinformatics* 2019;**20**:479.
55. Zhang W, Chen Y, Liu F, et al. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics* 2017;**18**:18.
56. Zhang Y, Zheng W, Lin H, et al. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics* 2018;**34**(5):828–35.
57. Zhang Y, Lin H, Yang Z, et al. A hybrid model based on neural networks for biomedical relation extraction. *J Biomed Inform* 2018;**81**:83–92.
58. Zhao Z, Yang Z, Luo L, et al. Drug-drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* 2016;**32**(22):3444–53.
59. Zheng W, Lin HF, Luo L, et al. An attention-based effective neural model for drug-drug interactions extraction. *BMC Bioinformatics* 2017;**18**:445.
60. Zhou D, Miao L, He Y. Position-aware deep multi-task learning for drug-drug interaction extraction. *Artif Intell Med* 2018;**87**:1–8.
61. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**(2018):i457–66.