The assessment of similarity vectors of fingerprint and UMLS in adverse drug reaction prediction

Milad Beshartifard[1], Zahra Ghorbanali[1], Fatemeh Zare-Mirakabad[1*]

[1]*Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran.*

**Corresponding Author: f.zare@aut.ac.ir*

Abstract

Identifying and controlling adverse drug reactions is a complex problem in the pharmacological field. Despite the studies done in different laboratory stages, some adverse drug reactions are recognized after being released, such as Rosiglitazone. Due to such experiences, pharmacists are now more interested in using computational methods to predict adverse drug reactions. In computational methods, finding and representing appropriate drug and adverse reaction features are one of the most critical challenges. Here, we assess fingerprint and target as drug features; and phenotype and unified medical language system as adverse reaction features to predict adverse drug reaction. Meanwhile, we show that drug and adverse reaction features represented by similarity vectors can improve adverse drug prediction. In this regard, we propose four frameworks. Two frameworks are based on random forest classification and neural networks as machine learning methods called F_RF and F_NN, respectively. The rest of them improve two state-of-art matrix factorization models, CS and TMF, by considering target as a drug feature and phenotype as an adverse reaction feature. However, machine learning frameworks with fewer drug and adverse reaction features are more accurate than matrix factorization frameworks. In addition, the F_RF framework performs significantly better than F_NN with ACC = %89.15, AUC = %96.14 and AUPRC = %92.9. Next, we contrast F_RF with some well-known models designed based on similarity vectors of drug and adverse reaction features. Unlike other methods, we do not remove rare reactions from the data set in our frameworks. The data and implementation of proposed frameworks are available at http://bioinformatics.aut.ac.ir/ADRP-ML-NMF/.

Keywords: Drug adverse reaction, Machine learning, Matrix factorization

1. Introduction

After the outbreak of coronavirus (SARS-CoV-2), according to studies and performed tests, the World Health Organization (WHO) issued an emergency use authorization for the drug hydroxychloroquine, which was canceled shortly afterward[1]. One of the reasons was the presence of a rare adverse reaction for this drug that caused heart disorders (cardiotoxicity)[2][3]. However, after its widespread use, this drug also caused many deaths[2]. In addition, it has been reported that about 26% of the people are admitted, only in a single

hospital in southern India, due to adverse drug reactions (ADRs)[4]. Moreover, drug toxicity is common among children[5] caused to be hospitalized in about 300 children with an average age of 5 years in a medical center in the Netherlands due to ADRs. Such problems show the importance of various drug assessments before a drug is produced and released to the market, considering its time-consuming and costly process.

Nevertheless, sometimes monitoring a drug after launch shows some rare adverse reactions, and it is caused to be withdrawn after a few years, e.g., Rosiglitazone[6]. In May 2007, after examining data from a clinic, it was found that taking Rosiglitazone had a significant effect on the deaths caused by cardiovascular diseases. Eventually, the drug was first withdrawn in Europe, and then in the same year, severe restrictions were imposed on its use in the United States[7].

Regardless of laboratory studies performed at various stages of drug production to identify its adverse reactions, this strategy is still not effective enough to solve the ADR problem. Therefore, there is a severe need to diagnose ADRs accurately. In this regard, researchers are interested in approaching the ADR problem by computational methods. Nowadays, recommender systems[8] and machine learning methods[9][10] have been common computational models for ADR prediction.

A recommender system can predict whether a user prefers an item based on its profile[11]. This technique is also used in the ADR problem. Drugs and adverse reactions are assumed as users and items, respectively. In other words, the recommender system predicts whether a drug has an adverse reaction based on the drug profile[12]. Galeano[13] introduced a collaborative filtering recommendation system to predict an adverse reaction for a new drug using known similar drugs. Matrix factorization[14] is a class of collaborative filtering recommendation systems. Poleksic et al.[15] used the compressed sensing (CS) model as a matrix factorization to predict unknown relationships between drugs and adverse reactions. In addition, the CS model is an appropriate model for dealing with sparsity data. It is suitable for solving the ADR problem because known drug-adverse reactions (positive data) are less than unknowns. In this method, the latent preference of drugs and adverse reactions is computed in a lower dimension by minimizing the defined loss function to complete the drug-adverse reaction associations. Also, later in 2020, Guo et al.[12] recovered drug- adverse reaction matrix using triple matrix factorization (TMF) model based on calculating the similarity between drugs and adverse reactions with different features.

In addition to recommender systems, various machine learning methods are also very effective in solving this problem. Chen et al.[16] predicted the possible likelihood of that drug being associated with adverse effects for each drug. In this method, the similarity of two drugs is

calculated based on the interaction of each drug with other drugs and target proteins. Finally, this algorithm uses the other drugs with this adverse reaction to calculate the score of a drug-adverse reaction association based on the relationship between the drugs and calculating the likelihood.

Khan[17] applied different learning models for predicting ADRs such as neural networks, support vector machine, random forest, naive Bayes, using different drug features like fingerprint and drug indications. He limited the SIDER database[18] based on ten adverse reactions with the maximum variance across the drugs. The negative and positive data are defined using the frequency of causing each side effect by a drug according to the recorded medical history of 30 patients in the SIDER database. Therefore, if the frequency of each drug-adverse reaction is more than 0.5, its association is considered positive and vice versa. Finally, for each adverse reaction, it is created a classification model. Zhao et al.[19] proposed a different approach based on the similarity of drugs with different properties such as fingerprints, the two-dimensional structure of drugs, target proteins, ATC code, and some features from the STITCH database. Similarity vectors are considered as the input of the random forest classifier model. Also, the positive data is determined based on known adverse reactions of the drugs with label 1. The negative data is randomly selected from unknown adverse reactions of drugs with label 0. In addition, they chose adverse reactions that are associated with more than five drugs. Rodriguez et al.[20] proposed a Bayesian network approach to predict ADRs using 593 pharmaceutical care center reports. Dey[21] introduced a model to convert two-dimensional or three-dimensional drug structures into numerical vectors using convolutional neural networks (CNNs) for each adverse reaction. Zheng et al.[22] calculated the similarity of drugs based on various properties such as chemical structure, target protein, drug alternatives, and ATC code to form the feature vectors of each drug-adverse reaction association as the model input. Also, they introduced a new approach for selecting negative data. The negative data is chosen based on the assumption that dissimilar drugs have fewer common adverse reactions. Uner[23] proposed a learning method to solve the ADR problem based on CNNs using the structural features of the drug and the characteristics of gene expression. They also selected negative data from pairs of drugs and adverse reactions whose relationship is unknown. In 2020, Liang et al. proposed a new approach for making negative data by random walk on drug-drug interaction networks[24]. They removed adverse reactions associated with less than six drugs from the dataset. In 2021, It was shown that combining different data sources can improve the accuracy of learning models for the ADR problem[25]. Zhang et al.[26] defined negative data based on drug-indication associations and applied a machine learning model to classify adverse reactions as adverse or therapeutic. Table 1 shows a summary of the studies on the ADR prediction problem.

References/Year	Drug Feature	Adverse reaction Feature	Algorithm	Number of Drugs	Number of adverse reactions	Database for adverse reactions	Database for Drug
2013[16]	Target protein Literature-Association	-	Machine Learning	835	100	-	SIDER STITCH
2017[17]	Fingerprint Indication Target Protein	-	Machine Learning	667	10	-	SIDER CHEMBL
2018[15]	Fingerprint	UMLS semantic	Recommender system	1430	5868	SIDER MedDRA	SIDER PubChem
2018[19]	Fingerprint Smiles ATC code Target Protein Literature association	-	Machine Learning	841	824	-	SIDER KEGG RDKit
2018[20]	ADR report	causality categories	Machine Learning	-	2100	Northern Pharmacovigilance Centr	-
2019[22]	Fingerprint Target Protein ATC code Substituent	-	Machine Learning	917	500	-	SIDER DrugBank EMBL-EBI
2019[23]	gene expression SMILES	-	Machine Learning	791	1052	-	SIDER PubChem
2020[27]	Fingerprint Side effect profile	Drug profile	Recommender System	614	5596	SIDER PubChem	SIDER
2020[24]	Fingerprint SMILES ATC code Target Protein Literature Association	-	Machine Learning	841	824	-	SIDER STITCH PubChem KEGG
2021[26]	Indication Target protein	CUI code	Machine Learning	3632	5589	SIDER -DrugBank	SIDER

Table 1:Some studies on the ADR problem.

In the computational methods, finding and representing appropriate features of drugs and adverse reactions are challenging. Here, we assess fingerprint and target as drug features; and phenotype and Unified Medical Language System (UMLS) as adverse reaction features to predict adverse drug reaction. Meanwhile, we show that drug and adverse reaction features represented by similarity vectors can improve the performance of the computational methods in adverse drug prediction.

This article proposes four frameworks to analyze drug features and adverse reaction features in drug-adverse reaction association prediction. Two frameworks are based on two machine learning methods; a random forest classifier[10] and a neural network[9]. The rest improve two well-known matrix factorization models; CS [14] and TMF[12].

The first framework is F_RF to predict drug-adverse reactions based on a random forest classifier by considering drug-drug similarity vectors and the vector of adverse reaction

similarity. F_RF approach is compared with the neural network-based framework, F_NN, to address the ADR problem. F_RF framework shows better performance than F_NN. Although we improve the CS model[15] into CS^{Phen} and TMF method[12] into three versions, TMF_{Targ}, TMF^{Phen} and TMF_{Targ}^{Phen} as matrix factorization approaches, the result announces that the F_RF model has better accuracy than state-of-the-art matrix factorization methods. Therefore, we compare the F_RF framework with some well-known algorithms in machine learning approaches that consider similarity vectors as features.

It seems that defining negative data similar to Zheng's approach[22] and considering similarity vectors as drug and adverse reaction features improve the performance of F_RF. Moreover, unlike some methods[24][19] that remove rare adverse-drug reactions, we consider all drugs and adverse reactions, including rare associations, as training data.

Although the limited association of rare adverse reactions can reduce the model's accuracy, the F_RF framework achieves comparable performance. Finally, our framework successfully predicts some associations in some case studies.

This paper is organized as follows: the "Materials and methods" section presents the databases and datasets, the notations and definitions, and a description of our proposed frameworks. The "Results" section includes assessing our frameworks and comparing results with other models. The "Discussion" section shows that the F_RF framework successfully predicts adverse reactions for some case study drugs, and finally, the "Conclusion" represents the future point of the ADR problem.

2. Material and methods

This section introduces drug and adverse reaction datasets and databases and then defines some notations to describe the problem of the adverse drug reaction (ADR). Finally, the proposed model for ADR prediction is explained in more detail.

2.1. Datasets and Databases

In the ADR problem, we need to extract drug and adverse reaction features in addition to known drug-adverse reaction associations from databases. In this paper, we generate three datasets called Δ_1, Δ_2, and Δ_3 from the SIDER database[18], and apply Poleksic[15] and Mizutani[28] datasets as Δ_4 and Δ_5, respectively.

We use the SIDER database[18] for collecting adverse reactions (CUI codes), drugs (CID codes), and drug-adverse reaction associations. We split the extracted adverse reactions and drugs from the SIDER database into three datasets, Δ_1, Δ_2 and Δ_3. Fig. 1. shows that the first dataset includes more common adverse reactions in most drugs. In Fig. 2., it can be seen Δ_2

dataset includes adverse reactions, which are not more common in most drugs. Fig. 3. indicates Δ_3 dataset contains rare adverse reactions.

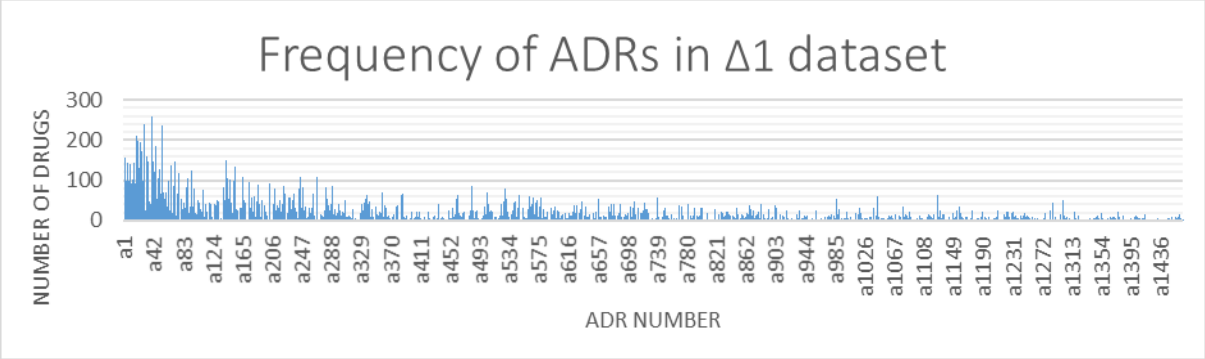

Fig. 1. Frequency of ADRs in Δ1 dataset.

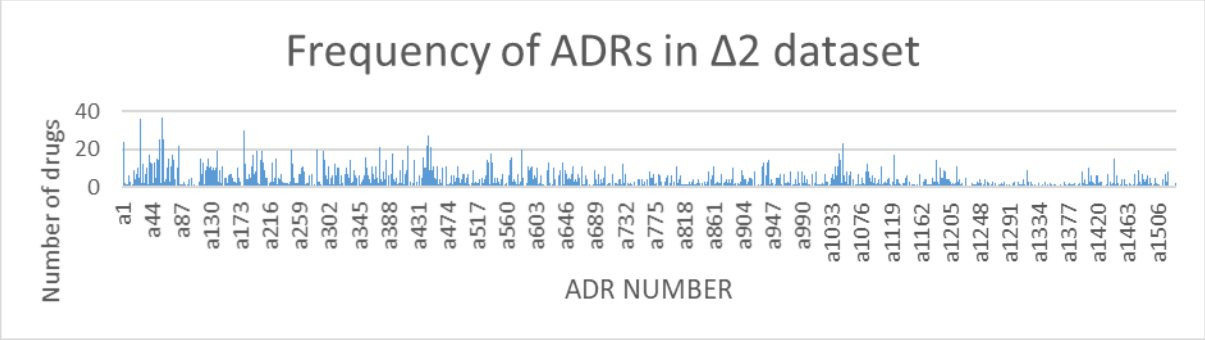

Fig. 2. Frequency of ADRs in Δ2 dataset.

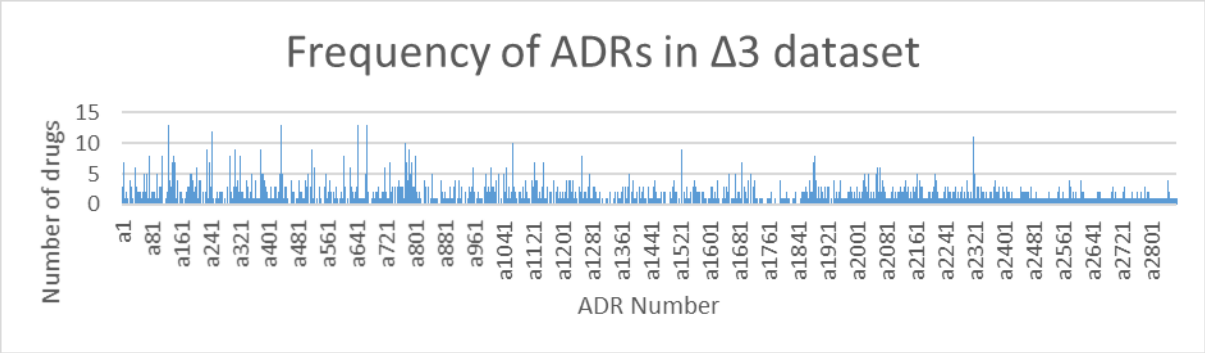

Fig. 3. Frequency of ADRs in Δ3 dataset.

Table 2 shows that dataset $\Delta \in \{\Delta_1, \Delta_2, \Delta_3\}$ includes a 2-tuple $\Delta = < D, A >$ where D and A indicate the sets of the selected drugs and adverse reactions, respectively. The numbers of drugs and adverse reactions are shown by $|D| = m$ and $|A| = n$, respectively. The first column shows three datasets. The second and third ones display the number of extracted drugs and adverse reactions, respectively, where each adverse reaction has been associated with at least one drug. The fourth column indicates the number of known drug-adverse reaction associations in each dataset. The next column presents the number of adverse reactions that are treated by some extracted drugs. We call treated adverse reactions as indications. Therefore, the set of indications is a subset of A as an adverse reaction set. The last column shows the number of drug-indication associations in each dataset.

Moreover, we apply Poleksic[15] and Mizutani[28] datasets as Δ_4 and Δ_5, respectively. Table 3 shows more details of these datasets. Meanwhile, we extract 17843 and 17418 phenotypes for the adverse reactions in Δ_4 and Δ_5 datasets from the CTD database, respectively.

The fingerprint of each drug as a binary vector with length 881 is extracted from PubChem[25] database. The directly interacted proteins with drugs of the Mizutani database are extracted from DrugBank[29] and Matador[30] databases. The number of these target proteins is equal to 1368.

$\Delta =< D, A >$	$\|D\| = m$	$\|A\| = n$	The number of drug-adverse reaction associations	The number of indications	The number of drug-indication associations
Δ_1	357	1467	24239	663	1498
Δ_2	443	1533	4591	955	3788
Δ_3	630	2868	3017	740	2786

Table 2:Extracted datasets from the SIDER database.

$\Delta =< D, A >$	$\|D\| = m$	|A|=n	The number of drug-adverse reaction associations
Δ_4	1430	5868	120491
Δ_5	658	1339	49051

Table 3: Datasets used in the Poleksic and Mizutani studies.

2.2. Notations and definitions

This part describes the selected biological features of a drug and an adverse reaction. Moreover, we define some notations for the feature representations.

2.2.1. Drug

A set of m drugs is denoted by $D = \{d_1, d_2, \dots, d_m\}$, where $d_i \in D$ shows the i^{th} drug. Each drug $d \in D$ is displayed by fingerprint chemical structure or protein targets as follows:

- The binary vector $F^d = [f_1, \dots, f_{881}]$ with length 881 represents fingerprint[31]. Each f_i with value 1 or 0 represents the existence or absence of the i^{th} substructure descriptor associated with a specific chemical feature, respectively.
- The binary vector $\Upsilon^d = [\tau_1, \tau_2, \dots, \tau_{1368}]$ with length 1368 shows target proteins. Each τ_i with value 1 or 0, represents the i^{th} protein as a known target for drug d or not, respectively.

To calculate the fingerprint chemical structure similarity of drugs $d, d' \in D$, we use Gaussian Interaction Profile (GIP) meter[12][32] defined as follows:

$$GIP^{Fing}(d, d') = \exp\left(-\gamma\left(\left|\left|F^{d} - F^{d'}\right|\right|\right)^{2}\right), \quad (1)$$

where the bandwidth control parameter (γ) is assigned one[12].

To measure the target protein similarity of drugs $d, d' \in D$, we apply the cosine similarity (COS) criterion[12], as follows:

$$COS^{Targ}(d, d') = \frac{\Upsilon^{d}(\Upsilon^{d'})^{T}}{\left|\left|\Upsilon^{d}\right|\right| \left|\left|\Upsilon^{d'}\right|\right|}. \quad (2)$$

For each drug $d \in D$ and similarity function $f \in \{GIP^{Fing}, COS^{Targ}\}$, we define the similarity vector $\delta^{d,f}$ with length m, as follows:

$$\delta^{d,f} = [f(d, d_1), f(d, d_2), \ldots, f(d, d_m)], \quad d_i \in D, \quad 1 \le i \le m, \quad (3)$$

where $\delta^{d,f}$ shows the feature representation of drug d.

2.2.2. Adverse reaction

The set of n adverse reactions is denoted by $A = \{a_1, a_2, \ldots, a_n\}$, where $a_j \in A$ shows the j^{th} adverse reaction. For each $a \in A$, phenotype or UMLS[33][15] is considered as an adverse reaction feature:

- The phenotype of adverse reaction a is shown by the binary vector $P^{a} = [p_1, \ldots., p_{18058}]$. In our datasets, 18058 is the union of all extracted phenotypes and p_k is considered 1 or 0 to show the existence and absence of the k^{th} phenotype for adverse reaction a. The phenotype illustrates each adverse reaction based on genetic ontology through biological processes, cellular components, and molecular functions[34].
- The UMLS[33] includes over 100 medical terminologies with a unified and semantic network designed by the National Library of Medicine to support scientific research.

To calculate the phenotype similarity of adverse reactions, $a, a' \in A$, we use the cosine similarity (COS) criterion as follows:

$$COS^{Phen}(a, a') = \frac{P^{a}(P^{a'})^{T}}{\left|\left|P^{a}\right|\right| \left|\left|P^{a'}\right|\right|}, \quad (4)$$

The UMLS similarity of adverse reactions $a, a' \in A$ is computed using UMLS-similarity software which is denoted by function $SIM^{UMLS}(a, a')$[33][15].

For each adverse reaction $a \in A$ and similarity function $f \in \{COS^{Phen}, SIM^{UMLS}\}$, we define the similarity vector $\alpha^{a,f}$ with length n, as follows:

$$\alpha^{a,f} = [f(a, a_1), f(a, a_2), \ldots, f(a, a_n)], \quad a_i \in A, \quad 1 \le i \le n, \quad (5)$$

where $\alpha^{a,f}$ indicates the feature vector of adverse reaction a.

2.3. Adverse drug reaction problem

We assume that $D = \{d_1, d_2, \dots, d_m\}$ and $A = \{a_1, a_2, \dots, a_n\}$ represent m drugs and n adverse reactions. In the ADR problem, biological features of drug $d \in D$ and adverse reaction $a \in A$ are given to the model. The primary goal of the ADR problem is to predict the association between adverse reaction $a \in A$ and drug $d \in D$. If the model predicts adverse reaction a associated with drug d, the output is one and otherwise zero.

For each drug d, we use the similarity vector $\delta^{d,GIP^{Fing}}$, or $\delta^{d,COS^{Targ}}$ as a drug feature. For each adverse reaction a, the similarity vector $\alpha^{a,COS^{Phen}}$ or $\alpha^{a,SIM^{UMLS}}$ is considered as an adverse reaction feature.

2-4. Proposed Models

In this paper, four frameworks are proposed to solve the ADR problem. As machine learning approaches, the first and second frameworks use random forest[10] and neural networks [9], respectively. The third and fourth ones improve CS[15] and TMF[12] as matrix factorization models for ADR prediction. In the following, we describe these frameworks in more detail.

The first and second frameworks are called F_RF and F_NN, respectively. In both models, the concatenation of the similarity vectors $\delta^{d,GIP^{Fing}}$ and $\alpha^{a,SIM^{UMLS}}$ is given as input to predict drug-adverse reaction association. For each drug d, the similarity vectors $\delta^{d,GIP^{Fing}}$ is computed using the GIP function (see Eq.1). Meanwhile, the similarity vector $\alpha^{a,SIM^{UMLS}}$ is obtained by the UMLS function[33] for each adverse reaction a. In F_RF and F_NN frameworks, we require positive and negative samples for training the model. Known drug-adverse reaction associations and known drug-indication associations are considered positive and negative data, respectively (see Fig. 4.).

The third framework named CS^{Phen} improves CS[15] model as a matrix factorization approach for ADR prediction. The original CS model uses UMLS-similarity software to calculate adverse reaction similarity for each $a, a' \in A$ shown by $SIM^{UMLS}(a, a')$. We improve the similarity matrix between adverse reactions in the CS model by adding the phenotype similarity of adverse reactions, COS^{Phen} (see Eq.4), to UMLS-similarity as follows:

$$\frac{SIM^{UMLS}(a, a') + COS^{Phen}(a, a')}{2}. \qquad (6)$$

The fourth framework is defined based on TMF[12] model as a matrix factorization approach for ADR prediction. This model combines different criteria to compute the similarity between two drugs $d, d' \in$. We call this similarity $S^{TMF}(d, d')$. Meanwhile, the original TMF uses drug

profile as a feature to compute the similarity between two adverse reactions $a, a' \in A$ named $S^{TMF}(a, a')$. Here, we define three different versions of TMF as follows:

- TMF_{Targ}: The drug-drug similarity matrix of TMF is changed as:

$$\frac{S^{TMF}(d, d') + COS^{Targ}(d, d')}{2}, \tag{7}$$

where $COS^{Targ}(d, d')$ is obtained based on Eq. 2 to show the similarity between targets of drugs.

- TMF^{Phen}: The similarity matrix between adverse reactions is developed as:

$$\frac{S^{TMF}(a, a') + COS^{Phen}(a, a')}{2}, \tag{8}$$

where COS^{Phen} is obtained based on Eq.4 to show the similarity between phenotypes of adverse reactions.

- TMF_{Targ}^{Phen}: Both drug-drug similarity matrix and the similarity matrix between adverse reactions are improved by Eq.7 and Eq.8.

The matrix factorization models define positive and negative data based on known drug-adverse reaction associations and unknown drug-adverse reaction associations, respectively.

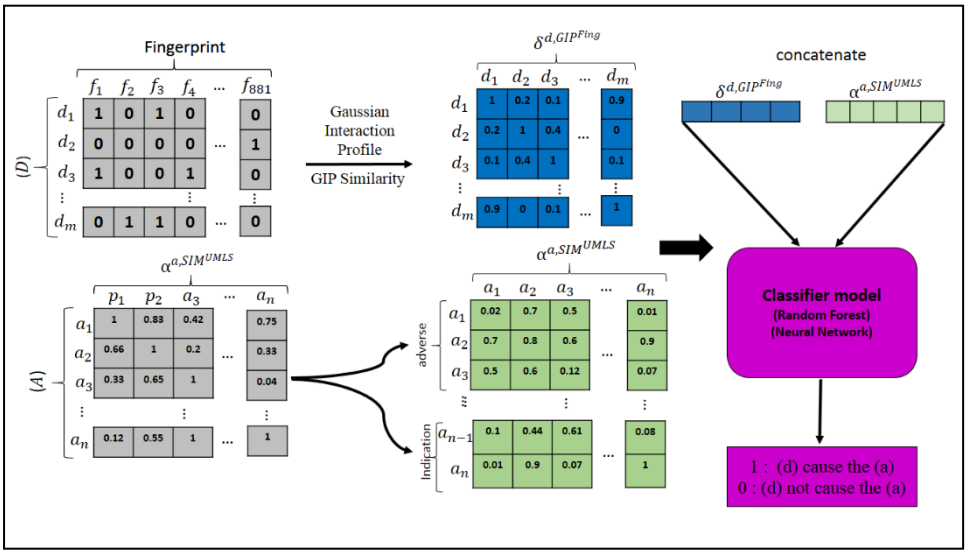

Fig. 4. Outline of the first and second frameworks.

3. Result

In this section, we evaluate the four proposed frameworks. The first and second models, F_RF and F_NN, are based on random forest classifier and neural network as machine learning methods. The third framework, CS^{Phen}, improvs CS model as a matrix factorization approach. In the fourth framework, we define three versions of the TMF model, TMF_{Targ}, TMF^{Phen} and

TMF_{Targ}^{Phen} , as the matrix factorization model. Each framework was implemented in Matlab 2018b under Windows and Intel Core i5-2430M processor and 4GB of memory.

In the following, we introduce our selected evaluation criteria, then the parameters of each framework are explained. Next, we assess the performance of machine learning frameworks F_RF and F_NN and analyze their effectiveness to predict rare adverse reactions. Then, the assessment of frameworks 3-4, based on matrix factorization, is evaluated. Later, we compare our proposed frameworks to introduce the best one and determine its performance against four well-known models. Finally, we assess our best framework effectiveness on predicting associations of some case studies.

3.1. Evaluation criteria

We evaluate our frameworks using the area under receiver operating characteristic curve (AUC), the area under precision-recall curve (AUPRC), and accuracy (ACC) criteria.

AUC is obtained based on the false positive rate (FPR) and the classifier model's real positive rate (TPR) under different classification thresholds, where:

$$FPR = \frac{FP}{FP + TN}, \qquad TPR = \frac{TP}{TP + FN},$$

and FP, TN, TP, and FN display false positive, true negative, true positive, and false negative, respectively (see Table 4).

Prediction	Definition
True Positive (TP)	the number of known drug-adverse reaction associations predicted correctly by the model
False Positive (FP)	the number of drug-indication associations predicted wrongly by the model
True Negative (TN)	the number of drug-indication associations predicted correctly by the model
False Negative (FN)	the number of known drug-adverse reaction associations predicted wrongly by the model

Table 4: Definition of true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

AUPRC shows the relationship between sensitivity (recall) and positive predictive value (precision), where:

$$precision = \frac{TP}{TP + FP}. \quad recall = \frac{TP}{TP + FN}.$$

ACC indicates the rate of correct prediction to all predictions as below:

$$ACC = \frac{TP + TN}{TP + FP + TP + FN}.$$

3.2. Parameters of the frameworks

The hyperparameters of F_RF, F_NN, CS^{Phen}, TMF_{Targ}, TMF^{Phen} and TMF^{Phen}_{Targ} models, respectively, are defined as follows:

- The F_RF framework

In Matlab 2018, the random forest classifier is located in the package Statistics and Machine Learning Toolbox[1] and has some hyperparameters which can be changed according to the problem. Here, we refer to the three most important ones:

  - "MinLeafSize" shows the minimum observations (samples) per leaf, which is essential in dividing the nodes in the decision trees. By default, this parameter is 1 for classification. A smaller number of "MinLeafSize" makes the model more prone to capturing noise in the training data.
  - "NumPredictorsToSample" means the number of predictor or feature variables to select at random for each decision split. By default, it is equal to the square root of the total number of variables for classification.
  - "NumLearningCycles" variable represents the number of decision trees in the random forest.

As it can be seen in Table 5, we find the best values for these parameters on each dataset $\Delta \in \{\Delta_1, \Delta_2, \Delta_3\}$ by trial and error.

Datasets	**Parameters**			**Evaluation criteria**		
	MinLeafSize	NumPredictorsToSample	NumLearningCycles	AUC	AUPRC	ACC
$\mathbf{\Delta_1}$	1	6	38	0.5054	0.3201	0.7301
	1	**3**	**600**	**0.9637**	**0.9417**	**0.9031**
	1	**6**	**38**	**0.9754**	**0.9354**	**0.9288**
	3	6	38	0.9601	0.9440	0.9050
$\mathbf{\Delta_2}$	3	6	11	0.9673	0.8292	0.9109
	1	9	11	0.9540	0.8611	0.9077
	1	6	42	0.9661	0.9255	0.9132
$\mathbf{\Delta_3}$	**1**	**6**	**38**	**0.9395**	**0.9055**	**0.8799**

Table 5: Hyperparameters used in the Random Forest model.

- The F_NN framework

We use a neural network strategy using the back-propagation approach for learning to predict drug- adverse reaction associations. Our model contains an input layer (according to the size of features) followed by one fully connected hidden layer (with100 neurons) and an output layer

[1] https://www.mathworks.com/help/stats/treebagger.html

that decides whether a drug and an adverse reaction are associated or not based on sigmoid function. The learning rate is 0.05, and we train the network about 100 iterations. Our activation function for the hidden layer is tan-sigmoid, and errors are backpropagated according to Scaled Conjugate Gradient (SCG) strategy.

- Improved CS and TMF models

The parameters of these two models are not changed and set according to the original models[12][15].

3.3. The assessment of machine learning methods on ADR problem

In this subsection, we assess F_RF and F_NN frameworks on each dataset $\Delta \in \{\Delta_1, \Delta_2, \Delta_3\}$. Moreover, we compare our models to find which one is more accurate on rare adverse reactions.

3.3.1. The assessment of F_RF and F_NN frameworks on Δ_1, Δ_2 and Δ_3 datasets

For each dataset $\Delta \in \{\Delta_1, \Delta_2, \Delta_3\}$ (see Table 2), we consider $\Delta = < D, A >$ where D and A indicate drug and adverse reaction sets. All known drug-adverse reaction associations are considered as positive data (P^{Δ}). Similar to Zhang et al.[26], we suppose that if a drug is prescribed for one of the adverse reactions, it does not cause that adverse reaction. According to this assumption, we extract drug-indication associations as negative data (N^{Δ}).

To form a test set from the dataset, we randomly select 10% of negative data (N^{Δ}_{test}) and the exact size of positive data (P^{Δ}_{test}). The rest of the positive ($P^{\Delta}_{train} = P^{\Delta} - P^{\Delta}_{test}$) and negative data ($N^{\Delta}_{train} = N^{\Delta} - N^{\Delta}_{test}$) are considered as the training data (see Table 6).

Datasets	**Associations in dataset**		**Associations in training data**		**Associations in test data**	
	$\lvert P^{\Delta}\rvert$	$\lvert N^{\Delta}\rvert$	$\lvert P^{\Delta}_{train}\rvert$	$\lvert N^{\Delta}_{train}$	$\lvert P^{\Delta}_{test}\rvert$	$\lvert N^{\Delta}_{train}\rvert$
$\boldsymbol{\Delta_1}$	24239	1498	24089	1348	150	150
$\boldsymbol{\Delta_2}$	4591	3788	4212	3409	379	379
$\boldsymbol{\Delta_3}$	3017	2786	2738	2507	279	279

Table 6: the number of training and test data in each dataset.

As it can be seen in Table 6, the training data extracted from Δ_1 dataset is imbalanced because the number of negative data is less than positive data. We balance this training data by oversampling strategy to repeat the negative data.

For each dataset Δ, we train F_RF and F_NN frameworks on $P^{\Delta}_{train} \cup N^{\Delta}_{train}$ and test on $P^{\Delta}_{test} \cup N^{\Delta}_{test}$. Table 7 shows the values of corresponding evaluation criteria on the test set. Both frameworks have better performance on Δ_1 dataset because the number of training data in

this dataset is more than the others with common adverse reactions. The average accuracy on all evaluation criteria on F_RF is more than in the F_NN framework.

$F\_RF(P^{\Delta}_{test} \cup N^{\Delta}_{test})$	AUC	AUPRC	ACC	$F\_NN(P^{\Delta}_{test} \cup N^{\Delta}_{test})$	AUC	AUPRC	ACC
Δ_1	**0.9637**	**0.9401**	**0.9032**	Δ_1	**0.9735**	**0.9486**	**0.9132**
Δ_2	0.9612	0.9076	0.8879	Δ_2	0.9344	0.9247	0.8694
Δ_3	0.9592	0.9392	0.8835	Δ_3	0.9308	0.9050	0.8513
Average	0.9614	0.9290	0.8915	**Average**	0.9462	0.9261	0.8780

Table 7:The evaluation criteria of F_RF and F_NN frameworks on the test set.

3.3.2. The assessment of F_RF and F_NN frameworks on the rare adverse reactions

An adverse reaction associated with a maximum of two drugs (and a minimum of one drug) is considered as adverse reaction rare. Predicting rare adverse reactions is an obstacle because the known associations of them are too limited. Some studies[24][19] exclude the rare adverse reactions from their dataset to increase their performance. The number of drug–rare adverse reaction associations in each dataset Δ_1, Δ_2 and Δ_3 contain 582, 668, and 1797, respectively. Fig. 1., Fig. 2., and Fig. 3. depict the ratio of the adverse reaction numbers and their related drugs in each dataset $\Delta \in \{\Delta_1, \Delta_2, \Delta_3\}$. Each adverse reaction in datasets Δ_1, Δ_2, and Δ_3 is associated with average 16.52, 2.99, and 1.05 drugs, respectively.

To evaluate the accuracy of F_RF and F_NN frameworks on rare adverse reactions in Δ dataset, we choose positive test data (P^{Δ}_{test}) with 10% drug-adverse reaction associations known as rare adverse reactions. The sets of N^{Δ}_{test}, P^{Δ}_{train} and N^{Δ}_{train} are generated the same as the previous section. Table 8 illustrates the results for predicting rare adverse reactions on the test set. In both models, the performance of Δ_3 dataset is better than other ones because it has more rare adverse reactions than the other datasets. The results show that F_RF is generally more accurate than the F_NN framework.

$F\_RF(P^{\Delta}_{test} \cup N^{\Delta}_{test})$	AUC	AUPRC	ACC	$F\_NN(P^{\Delta}_{test} \cup N^{\Delta}_{test})$	AUC	AUPRC	ACC
Δ_1	0.8390	0.7465	0.7400	Δ_1	0.8733	0.8456	0.8212
Δ_2	0.9136	0.8833	0.8298	Δ_2	0.8702	0.8534	0.7902
Δ_3	**0.9345**	**0.9182**	**0.8428**	Δ_3	**0.9101**	**0.9015**	**0.8127**
Average	0.8957	0.8493	0.8042	**Average**	0.8845	0.8668	0.8080

Table 8: The evaluation of F_RF and F_NN frameworks on rare adverse reactions as the test set.

3.4. The assessment of matrix factorization methods on ADR problem

In this subsection, we assess the performance of the improved CS model[15] called CS^{Phen}, and TMF model[12] in three versions, TMF_{Targ}, TMF^{Phen} and TMF^{Phen}_{Targ}.

As mentioned above, these models are extended by adding the new drug and adverse reaction features to the original ones. The main version of these models are performed on datasets Δ_4[15] and Δ_5 [28], respectively. Therefore, we perform the improved CS model, CS^{Phen}, and three different versions of the TMF model, TMF_{Targ}, TMF^{Phen} and TMF^{Phen}_{Targ}, on these datasets.

In these models, positive and negative data are defined based on all known drug-adverse reaction associations and unknown drug-adverse reaction associations, respectively.

To evaluate improved models, we perform cross-validation similar to the original version of CS[15] and TMF[12]. Here, we divide our known drug-adverse reaction associations randomly into equal subsets. One of them is chosen randomly, and its associations are set as 0 in the drug-adverse reaction associations matrix, called the test set. Then, the model is trained by the remaining subsets. For prediction evaluation, the test set is added to the whole matrix as positive samples again. This process is repeated for every subset.

Table 9 depicts the evaluation scores for general CS and TMF models and their improved versions.

Model ($dataset$)	AUC	AUPR
$\boldsymbol{CS(\Delta_4)}$	0.9412	0.5059
$\boldsymbol{CS^{phen}(\Delta_4)}$	**0.9526**	**0.5517**
$\boldsymbol{TMF(\Delta_5)}$	0.9415	0.7071
$\boldsymbol{TMF^{phen}(\Delta_5)}$	0.9447	0.7093
$\boldsymbol{TMF_{targ}(\Delta_5)}$	0.9436	0.7100
$\boldsymbol{TMF^{Phen}_{targ}(\Delta_5)}$	**0.9479**	**0.7104**

Table 9:The evaluation of improved and original CS and TMF models.

3.5. Comparison machine learning and matrix factorization methods

This paper makes two similarity matrices of drugs and adverse reactions based on different features of drugs and adverse reactions. Meanwhile, we design four frameworks. Two of them are based on machine learning, F_RF and F_NN, and the others are based on matrix factorization, improvement of CS[15], and TMF[12].

In machine learning frameworks, the contention of each row of similarity matrices is given as input to F_RF and R_NN. In matrix factorization frameworks, the similarity matrices are integrated into the original similarity matrices of CS and TMF approaches.

The results show that although we improve the CS and TMF models as matrix factorization approaches, F_RF and R_NN represent better performance as machine learning approaches. In

addition, fewer features are considered for drug and adverse reactions to computing the similarity in machine learning frameworks (see Table 7 and Table 9).

According to corresponding results of applying F_RF and F_NN on datasets (see Table 7 and Table 8), random forest performs more accurately on datasets and predicts rare adverse reactions with higher performance.

3.6. Comparison with Related Studies

We compare the best-proposed framework, F_RF, with three machine learning models[24][19][22] that predict drug- adverse reaction associations using similarity-based methods. In addition, we compare F_RF with the logistic regression model introduced by Zhang in 2021[26]. This model defines negative data as ours by considering drug-indication associations for negative data[26]. Table 10 illustrates the values of evaluation criteria for each one.

It should be noted, the Liang et al. model[24] excluded all adverse reactions with less than six associations from its dataset, and it positively affected the performance. However, the performance of F_RF on datasets includes rare adverse reactions. In addition to considering rare adverse reactions, F_RF uses fewer features than the others while the results are competitive with the rest models.

Moreover, the performance of F_RF is more accurate than Zheng et al.[22], which performs a classifier for each adverse reaction separately.

Model	Drug feature	Adverse reaction feature	Number of Drugs	Number of adverse reactions	AUC	AUPRC
F_RF	Fingerprint	UMLS	357	1467	0.9637	0.9401
			443	1533	0.9612	0.9076
			630	2868	0.9592	0.9392
Zhao et al. [19]	ATC code, literature (STITCH), Target protein	Drug profile	841	824	0.8015	-
Zheng et al. [22]	Fingerprint, target protein, substituent, ATC code,	-	917	500	0.9086	0.5424
Liang et al. [24]	Fingerprint, Simcomp (2D structure), ATC code, literature (STITCH),	Drug profile	841	824	0.98	0.98

Zhang et al. [26]	Target Protein, Drug Bank ID	CUI code	3632	5589	0.87	-

Table 10: Comparison of the proposed model with other related studies.

4. Discussion

We also use several antiviral drugs out of our drug set to test the performance of the F_RF model. It should be noted, the association of these drug–adverse reactions are not available in the training set. We assume that positive association is determined based on the predicted probability of an association between a drug-adverse reaction which is more than 0.5. In addition, we check these pairs with F_NN, too.

For drug darunavir, F_RF predicts two adverse reactions Cardiac arrest, a type of heart disorder, and Hepatic steatosis, a type of liver disorder with the probability of association 0.98 and 0.61 (see Table 11). it means darunavir causes these adverse reactions. According to [35][36], the model correctly makes the prediction.

Moreover, F_RF suggests drug ribavirin and adverse reaction Liver disorder as a negative association. It is mentioned in [37] which this drug can treat Liver disorder. Table 11 depicts the case studies and the results of F_NN and F_RF. F_NN confirms the results of F_RF.

Drug	Adverse Reaction	Type	known Association	The prediction score of F_RF	The prediction score of F_NN
darunavir (CID: 213039)	Cardiac arrest (CUI: C0018790)	heart disorer	drug-adverse reaction (in[36])	0.98	0.84
darunavir (CID: 213039)	Hepatic steatosis (CUI: C2711227)	liver disordere	drug-adverse reaction (in SIDER database [35])	0.61	0.58
ribavirin (CID: 5064)	Unspecified liver disorder (CUI: C0023895)	Liver disorder	drug-indiction(to treat the Liver disorder in [37])	0.45	0.25

Table 11:The results of three case studies.

5. Conclusion

This study proposed a framework called F_RF based on a random forest classifier to predict drug-adverse reaction associations. For this aim, a similarity vector is suggested by the drug-drug similarity score and the adverse reactions similarity function as the drug and adverse reaction representations. As the performance of machine learning methods depends on the training data, similarly to Zhang et al.[26], the drug-adverse reactions and drug-indication are considered positive and negative data, respectively. Then, another framework was introduced

using a neural network named F_NN. Comparing the corresponding of these frameworks indicated the F_RF got higher evaluation scores than F_NN for predicting rare and non-rare adverse reactions. Later, two state-of-the-art matrix factorization methods, CS and TMF, were improved to CS^{Phen} and TMF_{Targ}, TMF^{Phen} and TMF_{Targ}^{Phen} and contrast with F_RF. According to the results, F_RF is performed more accurately than these models. Moreover, F_RF framework was compared with some popular machine learning approaches in ADR problem. Although some methods exclude rare adverse reactions [24][19] or use more features to solve the ADR problem, F_RF utilized all drug-adverse reactions, including rare ones, and fewer features. The results announced that the F_RF performance was better than most of them. Meanwhile, F_RF correctly predicted cardiac arrest and hepatic steatosis as darunavir adverse reactions and suggested ribavirin to treat the liver disorder.

We conclude that using similarity vectors as drug and adverse reaction features and considering drug indications as negative data can improve drug-adverse reaction association prediction. Moreover, applying a random forest classifier with less computational complexity than other models achieves higher performance scores.

In the future, we aim to assess the 3D structures of drugs to increase the performance of drug-adverse reaction association prediction. In addition, applying drug-related clinical information can improve the accuracy of the model.

Declaration of competing interest

The authors declare no conflicts of interest relevant to the manuscript contents.

References

[1] "Coronavirus (COVID-19) Update: FDA Revokes Emergency Use Authorization for Chloroquine and Hydroxychloroquine | FDA." https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-revokes-emergency-use-authorization-chloroquine-and (accessed Nov. 04, 2021).

[2] B. P *et al.*, "Hydroxychloroquine: a comprehensive review and its controversial role in coronavirus disease 2019," *Ann. Med.*, vol. 53, no. 1, pp. 117–134, Jan. 2021, doi: 10.1080/07853890.2020.1839959.

[3] H. Zhao, J. Wald, M. Palmer, and Y. Han, "Hydroxychloroquine-induced cardiomyopathy and heart failure in twins," *J. Thorac. Dis.*, vol. 10, no. 1, p. E70, Jan. 2018, doi: 10.21037/JTD.2017.12.66.

[4] M. Venkatasubbaiah, P. D. Reddy, and S. V. Satyanarayana, "Analysis of the Adverse Drug Reactions and Associated Cost Burden on the Patients in a South Indian Teaching Hospital," *J. Adv. Med. Med. Res.*, pp. 88–98, Nov. 2020, doi: 10.9734/JAMMR/2020/V32I2130697.

[5] A. T. M. Dittrich, J. M. T. Draaisma, E. P. van Puijenbroek, and D. M. W. M. Te Loo, "Analysis of Reporting Adverse Drug Reactions in Paediatric Patients in a University Hospital in the Netherlands," *Paediatr. Drugs*, vol. 22, no. 4, pp. 425–432, Aug. 2020, doi: 10.1007/S40272-020-00405-3.

[6] A. M, T. AM, H. N, and P. S, "Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: current status, challenges and the future," *Ther. Adv. drug Saf.*, vol. 11, 2020, doi: 10.1177/2042098620938595.

[7] W. AJ and G. KL, "Rosiglitazone: a review of its use in the management of type 2 diabetes mellitus," *Drugs*, vol. 62, no. 12, pp. 1805–1837, 2002, doi: 10.2165/00003495-200262120-00007.

[8] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative Filtering Recommender Systems," *Found. Trends® Human–Computer Interact.*, vol. 4, no. 2, pp. 81–173, 2011, doi: 10.1561/1100000009.

[9] R. Féraud and F. Clérot, "A methodology to explain neural network classification," *Neural Networks*, vol. 15, no. 2, pp. 237–246, Mar. 2002, doi: 10.1016/S0893-6080(01)00127-7.

[10] "Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. R News, 2, 18-22. - References - Scientific Research Publishing." https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1274692 (accessed Nov. 04, 2021).

[11] ResnickPaul and V. R., "Recommender systems," *Commun. ACM*, vol. 40, no. 3, pp. 56–58, Mar. 1997, doi: 10.1145/245108.245121.

[12] G. X, Z. W, Y. Y, D. Y, T. J, and G. F, "A Novel Triple Matrix Factorization Method for Detecting Drug-Side Effect Association Based on Kernel Target Alignment," *Biomed Res. Int.*, vol. 2020, 2020, doi: 10.1155/2020/4675395.

[13] D. Galeano and A. Paccanaro, "A Recommender System Approach for Predicting Drug Side Effects," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, Oct. 2018, doi: 10.1109/IJCNN.2018.8489025.

[14] L. H, G. P, X. L, and P. A, "Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem," *Sci. Rep.*, vol. 6, Dec. 2016, doi: 10.1038/SREP38860.

[15] A. Poleksic and L. Xie, "Predicting serious rare adverse reactions of novel chemicals," *Bioinformatics*, vol. 34, no. 16, pp. 2835–2842, Aug. 2018, doi: 10.1093/BIOINFORMATICS/BTY193.

[16] C. L *et al.*, "Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions.," *Biomed Res. Int.*, vol. 2013, pp. 485034–485034, Sep. 2013, doi:

10.1155/2013/485034.

[17] M. Khan, "Drug side-effect prediction using machine learning methods," 2017.

[18] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. Database issue, p. D1075, 2016, doi: 10.1093/NAR/GKV1075.

[19] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Math. Biosci.*, vol. 306, pp. 136–144, Dec. 2018, doi: 10.1016/J.MBS.2018.09.010.

[20] P. P. Rodrigues, D. Ferreira-Santos, A. Silva, J. Polónia, and I. Ribeiro-Vaz, "Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network," *Artif. Intell. Med.*, vol. 91, pp. 12–22, Sep. 2018, doi: 10.1016/J.ARTMED.2018.07.005.

[21] S. Dey, H. Luo, A. Fokoue, J. Hu, and P. Zhang, "Predicting adverse drug reactions through interpretable deep learning framework," *BMC Bioinformatics*, vol. 19, Dec. 2018, doi: 10.1186/S12859-018-2544-0.

[22] Y. Zheng, H. Peng, S. Ghosh, C. Lan, and J. Li, "Inverse similarity and reliable negative samples for drug side-effect prediction," *BMC Bioinforma. 2019 1913*, vol. 19, no. 13, pp. 91–104, Feb. 2019, doi: 10.1186/S12859-018-2563-X.

[23] O. C. Uner, R. G. Cinbis, O. Tastan, and A. E. Cicek, "DeepSide: A Deep Learning Framework for Drug Side Effect Prediction," *bioRxiv*, p. 843029, Nov. 2019, doi: 10.1101/843029.

[24] L. H, C. L, Z. X, and Z. X, "Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy," *Comput. Math. Methods Med.*, vol. 2020, 2020, doi: 10.1155/2020/1573543.

[25] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant, "PubChem as a public resource for drug discovery," *Drug Discov. Today*, vol. 15, no. 23–24, pp. 1052–1057, Dec. 2010, doi: 10.1016/J.DRUDIS.2010.10.003.

[26] F. Zhang, B. Sun, X. Diao, W. Zhao, and T. Shu, "Prediction of adverse drug reactions based on knowledge graph embedding," *BMC Med. Informatics Decis. Mak. 2021 211*, vol. 21, no. 1, pp. 1–11, Feb. 2021, doi: 10.1186/S12911-021-01402-3.

[27] S. Shabani-Mashcool, S. A. Marashi, and S. Gharaghani, "NDDSA: A network- and domain-based method for predicting drug-side effect associations," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102357, Nov. 2020, doi: 10.1016/J.IPM.2020.102357.

[28] M. S, P. E, S. V, G. S, and Y. Y, "Relating drug-protein interaction network with drug side effects," *Bioinformatics*, vol. 28, no. 18, Sep. 2012, doi: 10.1093/BIOINFORMATICS/BTS383.

[29] W. DS *et al.*, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, no. Database issue, Jan. 2008, doi: 10.1093/NAR/GKM958.

[30] S. Günther *et al.*, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic Acids Res.*, vol. 36, no. suppl_1, pp. D919–D922, Jan. 2008, doi: 10.1093/NAR/GKM862.

[31] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "PubChem: Integrated Platform of Small Molecules and Biological Activities," *Annu. Rep. Comput. Chem.*, vol. 4, pp. 217–241, Jan. 2008, doi: 10.1016/S1574-1400(08)00012-1.

[32] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, Nov. 2011, doi: 10.1093/BIOINFORMATICS/BTR500.

[33] M. BT, P. T, and P. SV, "UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity.," *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2009, pp. 431–435, Nov. 2009, Accessed: Nov. 04, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/20351894/?tool=EBI.

[34] D. AP *et al.*, "Comparative Toxicogenomics Database (CTD): update 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, Jan. 2021, doi: 10.1093/NAR/GKAA891.

[35] "Side effect information for Darunavir." http://sideeffects.embl.de/drugs/213039/ (accessed Nov. 04, 2021).

[36] Y. J *et al.*, "Pharmacotherapy Management for COVID-19 and Cardiac Safety: A Data Mining Approach for Pharmacovigilance Evidence from the FDA Adverse Event Reporting System (FAERS)," *Drugs - real world outcomes*, vol. 8, no. 2, pp. 131–140, Jun. 2021, doi: 10.1007/S40801-021-00229-8.

[37] P. Ljungman *et al.*, "Oral Ribavirin for Prevention of Severe Liver Disease Caused by Hepatitis C Virus During Allogeneic Bone Marrow Transplantation," *Clin. Infect. Dis.*, vol. 23, no. 1, pp. 167–169, Jul. 1996, doi: 10.1093/CLINIDS/23.1.167.