Identifying drug interactions using machine learning

Idris Demirsoy[1,A–C,F], Adnan Karaibrahimoglu[2,D,E]

[1] Department of Computer Engineering, Faculty of Engineering, Uşak University, Turkey
[2] Department of Biostatistics and Medical Informatics, Suleyman Demirel University, Isparta, Turkey

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation;
D – writing the article; E – critical revision of the article; F – final approval of the article

Advances in Clinical and Experimental Medicine, ISSN 1899–5276 (print), ISSN 2451–2680 (online) *Adv Clin Exp Med*. 2023;32(8):829–838

Address for correspondence
Idris Demirsoy
Email: idrisdemirsoy1@gmail.com

Funding sources
None declared

Conflict of interest
None declared

Acknowledgements
Idris Demirsoy would like to thank all the people whose assistance was a milestone in the completion of this project. The author thanks Mondira Bhattacharya for introducing him to people who are expert in their field, Jaishri Meer for providing detailed information about drug–drug interactions and showing him MedDRA system, among others. The author would like to thank Tim Carlson for explaining the clinical drug–drug interactions, among others, Erhan Berber for emphasizing patient view on drug–drug interactions, Balachandra Ambiga for selecting him for the project, and Shereen McIntyre for believing in him and supporting him.

Abstract

The majority of Americans, accounting for 51% of the population, take 2 or more drugs daily. Unfortunately, nearly 100,000 people die annually as a result of adverse drug reactions (ADRs), making it the 4th most common cause of mortality in the USA. Drug–drug interactions (DDIs) and their impact on patients represent critical challenges for the healthcare system. To reduce the incidence of ADRs, this study focuses on identifying DDIs using a machine-learning approach. Drug-related information was obtained from various free databases, including DrugBank, BioGRID and Comparative Toxicogenomics Database. Eight similarity matrices between drugs were created as covariates in the model in order to assess their influence on DDIs. Three distinct machine learning algorithms were considered, namely, logistic regression (LR), eXtreme Gradient Boosting (XGBoost) and neural network (NN). Our study examined 22 notable drugs and their interactions with 841 other drugs from DrugBank. The accuracy of the machine learning approaches ranged from 68% to 78%, while the F1 scores ranged from 78% to 83%. Our study indicates that enzyme and target similarity are the most significant parameters in identifying DDIs. Finally, our data-driven approach reveals that machine learning methods can accurately predict DDIs and provide additional insights in a timely and cost-effective manner.

Key words: prediction, machine learning algorithms, drug–drug interaction, similarity matrices, biostatistics

Received on January 18, 2023
Reviewed on April 3, 2023
Accepted on July 21, 2023

Published online on August 14, 2023

Cite as
Demirsoy I, Karaibrahimoglu A. Identifying drug interactions using machine learning. *Adv Clin Exp Med*. 2023;32(8):829–838. doi:10.17219/acem/169852

DOI
10.17219/acem/169852

Introduction

A drug interaction occurs when one drug affects another drug by increasing or decreasing its effect. The manipulated effect of the drug might cause unwanted and unexpected side effects. Important pharmacological cycles that impact bioavailability, including assimilation, dissemination, digestion, and discharge, may be influenced by communication.[1] Such associations can incorporate issues like the organization of a medication that raises digestive system motility and decreases the retention of other medications, as well as the rivalry for a similar plasma protein carrier, restraint of the activity of an administered medication, or a communication at the discharge level, which influences the end of one of the medications.[2] Additionally, pharmacodynamic interactions may take place at the pharmacological level when 2 drugs interact with the same protein, at the signal level involving different signaling pathways, or at the effector level where different pharmacological responses are elicited.[3] Hence, it is not feasible to identify or predict all drug interactions. However, it is possible to evaluate and comprehend the most important features affecting drug–drug interactions (DDIs) using the approach proposed in our study. Our method provides satisfactory results and insight into which covariates are most used while identifying DDIs.

Three types of approach are mainly implemented to predict DDI methods using machine learning approaches: 1) a similarity-based approach, 2) a classification-based approach and 3) a text mining approach. Some examples of these methods[4] use similarity-based modeling with few similarity matrices such as 2D molecular structure, interaction profile, target, and side-effect similarities. On the other hand,[5] some use a large-scale logistic regression (LR) model to predict potential DDIs. From the chemical–protein interactome (CPI) profile-based similarity to MeSH-based similarity, 10 similarities were used.[6] Wu et al. provided a 3-tiered hierarchical text mining approach for Drug-Drug Interaction (DDI) analysis, designed to label essential terms, sentences containing drug interactions, and pairs of interacting drugs.[6] However, text mining was not preferred by researchers, as they used classification-based approach[7] to apply a new trend of similarity-based methods. Using the idea proposed by Vilar et al., a weighted similarity network was developed.[2] Similarly, Rohani and Eslahchi used a feature matrix with only using one machine learning algorithm, a neural network (NN).[7] Most of these classification-based machine learning algorithms only use either a predictor or multiple predictors which are similar to other researches with LR. We offer to use more predictors with advanced machine learning algorithms.

Predicting DDIs can be investigated as a binary classification problem,[7,8] where the dependent variable is the interaction or non-interaction; the goal is to correctly label the DDIs. Our method uses feature matrices with classification-based machine learning algorithms, not only to compare different algorithms but also to write and create our dataset, as explained in Feature matrices section. In Materials and methods section, we pointed out where the data was collected and which databases were used. Additionally, we briefly introduced the machine learning algorithms used in the paper. Sample feature matrices are shown in Table 1. The first 2 columns stand for drug 1 and drug 2 names, others represent 8 feature matrices. Each of them, independently, shows how similar 2 drugs are based on a specific feature. Under the drug name, there is a code that stands for the DrugBank ID for that specific drug. It helps to work on drugs based on data structures such as split and merge. In Table 1, DB00176 stands for fluvoxamine which is a serotonin reuptake inhibitor used to treat obsessive-compulsive disorder. Fluvoxamine is also one of the most used antidepressants with cardiovascular drugs.[9] Therefore, it has been chosen for illustration purposes in Table 1. The drugs in the 2nd column were chosen randomly. For example,

Table 1. Sample feature matrices

Drug 1, DrugBank ID	Drug 2, DrugBank ID	Pathway	MedDRA	Molecular	ATC	Target	Enzyme	PPI	Disease	Y
Fluvoxamine DB00176	norethisterone DB00717	0.881	0.171	0.065	0.000	0.010	0.179	0.211	0.172	1
Fluvoxamine DB00176	candesartan DB13919	0.885	0.112	0.191	0.000	0.015	0.206	0.174	0.071	1
Fluvoxamine DB00176	ergotamine DB00696	0.894	0.145	0.141	0.122	0.018	0.318	0.129	0.237	1
Fluvoxamine DB00176	aldosterone DB04630	0.879	0.050	0.090	0.000	0.008	0.152	0.170	0.163	0
Fluvoxamine DB00176	norfloxacin DB001059	0.886	0.173	0.141	0.000	0.009	0.328	0.232	0.296	1
Fluvoxamine DB00176	ropivacaine DB00296	0.938	0.162	0.188	0.123	0.006	0.336	0.208	0.139	1

MedDRA – Medical Dictionary for Regulatory Activities; ATC – Anatomical Therapeutic Chemical; PPI – protein–protein interaction. The dataset contains 11 columns, the first 2 are drug names. The next 8 columns are feature matrices that show similarities between 2 drugs in the sense of the concept. The final column, Y, shows whether or not 2 drugs are interacting: 1 corresponds to interactions and 0 corresponds to non-interactions.

Table 2. Comparison of the methods based on evaluation metrics

Algorithms	Accuracy	Sensitivity	Specificity	F1 Score	Kappa
Logistic regression	0.6864	0.6837	0.6988	0.7816	0.2641
Neural network	0.7510	0.7501	0.7533	0.8137	0.4154
XGBoost	**0.7812**	0.7911	0.7613	**0.8310**	0.5508

XGBoost – eXtreme Gradient Boosting. The largest numbers are bolded.

DB00717 stands for norethisterone, a progesterone used for birth control. The table can read Pathway similarities between them as 0.881, Medical Dictionary for Regulatory Activities (MedDRA) (side effect) similarities as 0.171, ATC similarities as 0.000, enzyme similarities as 0.179, and so on. Feature matrices section explains the methods used to compute each column. In the Data analysis section, the methods of data analysis are described in detail. The data were fitted using 3 machine learning algorithms and the obtained minimum accuracy was 68.64%, the minimum F1 score was 78.16% from LR, the maximum accuracy was 78.12%, and the maximum F1 score was 83.1% from eXtreme Gradient Boosting (XGBoost) (Table 2). In Conclusions section, the importance of features is explained, and the most important covariate for explaining DDIs is plotted for trained data. Finally, future work is discussed.

Materials and methods

When a patient takes numerous medications, clinical DDIs can develop. Clinical toxicity or treatment failures can be caused by these DDIs. Therefore, DDI evaluations are an important element of medication development and the risk–benefit analysis of novel treatments. In their DDI guidance documents, regulatory agencies such as the Food and Drug Administration (FDA) of the USA, the Pharmaceuticals and Medical Devices Agency (PMDA) of Japan, and the European Medicines Agency (EMA) have recommended various methodologies (in vitro, clinical, and in silico) to examine DDI potentials, which can be utilized with patient management strategies. In this project, we conducted non-clinical DDI testing. Therefore, in this paper, publicly available databases were used. The following databases were used:

- DrugBank: A freely available database that stores drug information such as the Anatomical Therapeutic Chemical (ATC) codes enzymes, transporters and interactions;
- TwoSIDES: A freely available database with information about adverse drug reactions, side effects and drug indications;
- BioGRID: A public database that stores and disseminates data on genetic and protein interactions in human models and organisms;
- PubChem: A reference database for drug structures;
- UniProt: A database of protein sequences and functions that is open to the public.

Similarity metrics

Distance or similarity metrics were used in a broad range of applications, prompting evaluations of their effectiveness in fields such as texture image retrieval, web page clustering and social media event detection.[10] The 3 most common similarity metrics include the Tanimoto coefficient, the Dice coefficient and the Cosine similarity.

The similarity values obtained using these methods were non-negative and included values between 0 to 1, including 0 and 1. When there was no similarity, a value of 0 was selected, and for many similarities, a value of 1 was chosen. The formulas compute these metrics and more can be found in Bajusz et. al.[10] We mainly used Tanimoto similarity metrics while computing our feature matrices (Equation 1):

$$\text{Tanimoto coefficient} = T(A, B) = \frac{c}{a + b - c} \quad (1)$$

where a is the total number of objects in A, b is the total number of objects in B and c is the number of common objects between A and B, in which A and B stands for the features of drug 1 and drug 2.

Inverse document frequency

In the raw frequency, all terms are given equal importance.[11] However, it is known that some terms are repeated frequently, but they are not as important as once thought. Therefore, inverse document frequency (IDF) is a metric for determining whether a phrase is common or uncommon in a corpus of documents.[12] For example, after applying IDF for side effect (MedDRA) similarity, in the beginning, all of the reported side effects are counted, the number of unique side effects is identified, and a frequency table is created. Finally, the IDF of each unique side effect is computed using the following formula (Equation 2):

$$\text{IDF}(t, \text{drugs}) = \log\left[\frac{n}{df\,(t, \text{drugs})}\right] \quad (2)$$

where N is the total number of documents, t denotes the interest, and df(t, drugs) denotes the number of medicines with the interest. If a side effect is reported frequently, it gets a smaller IDF number because of the natural logarithm of the fraction.

XGBoost

The XGBoost is one of the most preferred classification methods. It is not only computationally fast, but it also gives accurate results when compared to some other algorithms.[13] The XGBoost is a tree-based algorithm that is similar to a decision tree; however, it uses a parallel computing feature. The XGBoost uses base/weak learners, which are only slightly better than guessing, but combines a bunch of the weak learners to create a strong learner, which is a form of ensemble learning. It weighs each weak learner prediction based on its prediction performance.[14] The XGBoost uses 3 main forms of gradient boosting inside the algorithm. The learning rate is contained in gradient boosting, also known as the gradient boosting machine. For the training, test and validation sets, stochastic gradient boosting operates as a random sub-sample at the row and column level (if applicable). Finally, Regularized Gradient Boosting contains both L1, also known as Lasso regularization, and L2, also known as Ridge regularization. The XGBoost was chosen for this study because of its speed and excellent overall performance. The XGBoost constructs an ensemble of classification trees (as in classification and regression tree (CART)). When adding the t^{th} tree to the ensemble, the objective function for XGBoost can be formulated as follows (Equation 3):

$$\mathrm{Obj}(\theta) = L(\theta) + \Omega(\theta) \tag{3}$$

where $L(\theta)$ denotes the loss function and $\Omega(\theta)$ denotes the regulation function. More details of XGBoost can be found in Chen et al.[15]

Neural network

An NN is a computer program that uses algorithms to find patterns in a group of data.[16] It resembles the work of human brain which tries to find relationships between things. In this context, a NN is a type of nervous system that can be biological or artificial.

A simple NN contains 3 layers: input, hidden and output. A layer is a collection of neurons. Usually, the number of hidden layers defines the name of the network. More details about NNs can be found in many valuable sources, such as the study by Paul and Singh.[17] Activation functions are a very important part of NNs since they play a critical role in learning and understanding non-linear relationships between the input and output signals. The activation function in a NN receives a signal from the input, transforms it into an output signal, and sends it to the next layer. Except for the output layer, the rectified linear unit (ReLU) is a piecewise linear function used as an activation function because it has constraints on weights. When the value is less than 0, it takes 0; otherwise, it takes the obtained value.[18] There are a few advantages of using ReLUs. Because of the constraint, taking the inverse of the function is easy. Moreover, ReLU does not activate all neurons, which helps in fast computation. For the output layer, sigmoid function has been chosen (Equation 4):

$$\mathrm{Sigmoid} = S(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

Sigmoid is mainly chosen because as an output we are interested in interaction (1) and non-interaction (0).[19] In Equation 4, e stands for Euler's number.

Logistic regression

Logistic regression is a statistical model in which dependent variables are categorical and have only 2 levels, such as success/failure and interaction/non-interaction. Principally, LR is suitable to test the hypothesis about relationships between a dichotomous response and one or more categorical or continuous explanatory variables.

Since LR can accommodate multiple-level categorical dependent variables, but we are interested in 2 levels only, we defined the basic LR form as follows (Equation 5):

$$\mathrm{logit}(Y) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \tag{5}$$

where π is the probability of the outcome of interest, β_0 is the intercept, β_1 is the slope, and X is an arbitrary explanatory variable.

The LR-fitted line plot has a sigmoid or S-shaped curve and misrepresents data when using the linear regression line. Hence, LR uses logit, which stands for natural logarithm, to respond to variables. More details on LR can be found in the study by Peng et al.[20]

Feature matrices

The DDI data were collected from different sources. DrugBank database v. 5.6. (https://go.drugbank.com; retrieved by June 2020) is a free drug information source that provides over 1.3 million accurate and updated DDIs covering all FDA- and Health Canada-approved drug sources from drug labels and references. Another free source, the Comparative Toxicogenomics Database (http://ctdbase.org/downloads/; retrieved June 2020) is used for pathway and disease similarity. The TwoSIDES database (http://tatonettilab.org/resources/nsides/; retrieved June 2020) includes the information on adverse effects of the interacting drugs.

The DDI dataset, which was downloaded from DrugBank, was transformed into a matrix with dichotomous values [0,1] representing the interaction between 2 drugs. The extra columns take values between 0 and 1, which shows the similarity between the 2 drugs based on specific features such as chemicals, enzymes and proteins. When it comes to predicting new interactions, our method provides DDI predictions using the similarity between known DDIs and new drug pairs. Our basic assumption is that if drug A has a known interaction with drug B and

the similarity between drug B and new drug C is greater than a certain threshold, then drug A will interact with drug C. If the 2 drugs are known to interact, and there is another drug that is comparable to one of the drugs in the DDI pair, the 3rd drug can cause a DDI.

Feature vectors

One of the most important aspects of any statistical learning method is the extraction of a meaningful collection of features. Classic prediction studies primarily consider topological features. Ding et al. points that machine learning-based approaches can be divided into 3 groups.[21] We combined all these types to create a feature matrix that contains approaches from all these types. In these types of setups, each feature matrix is an explanatory variable. Therefore, it would be beneficial to have as many feature matrices as possible. After an extensive literature review, we found out that creating the following feature matrices is feasible and advantageous.

Molecular structure similarity

To begin with, we present a specified collection of chemicals as simplified molecular-input line-entry system (SMILES) strings obtained from DrugBank. The RDKit was used to transform the SMILES strings into molecular extended connectivity fingerprints (ECFP; Open-Source Cheminformatics Software).[22] The two-dimensional Tanimoto similarity measure, also known as the Jaccard similarity measure of the fingerprints, was used to calculate the similarity scores between 2 drug molecules.

ATC similarity

The World Health Organization (WHO) uses the Anatomical Therapeutic Chemical (ATC) classification system, which is a hierarchical classification system that organizes medications by organ or system. The ATC codes were gathered from DrugBank. Rednik's semantic similarity algorithm was used to compute ATC similarity. This type of similarity was evaluated using ATC codes which are shown in Fig. 1. The ATC coding system partitions are based on the biological system or organ on which they target.

Target similarity

Drug targets, such as specific proteins and nucleic acids, are a type of biological macromolecule in the body that has a pharmacodynamic function through interacting with medicines. To predict drug–target interactions (DTIs), several researchers used a single similarity measure for medications and targets, namely chemical structure similarity for pharmaceuticals and amino acid sequence similarity for targets. Amino acid sequences of the target proteins were obtained from the Universal Protein (UniProt) database.

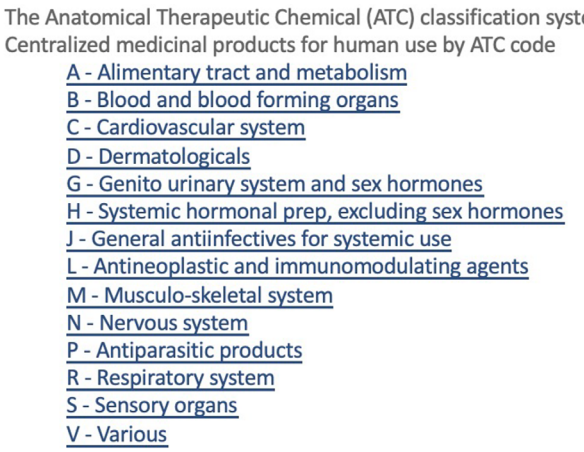

Fig. 1. The Anatomical Therapeutic Chemical (ATC) classification system

Then, the method suggested by Bleakley and Yamanishi[23] was used, namely the Smith–Waterman sequence alignment score between drug target genes computed with BLOSUM62 substitution matrix.[24] The scores' geometric means were computed to normalize the score, which was obtained after aligning each sequence (Fig. 2).

Pathway similarity

A drug's target proteins are found in a variety of pathways, which means that a single drug can affect numerous pathways and modify their activities. Pathway information on drugs was obtained from the Comparative Toxicogenomics Database.

We combined this information with DrugBank data. The IDF was used to assign more weight, and the Tanimoto coefficient was utilized to compute pathway similarities between pairs of drugs.

Disease similarity

Identifying drug–disease associations is time-consuming and expensive. Information on diseases associated with drugs was extracted from the Comparative Toxicogenomics Database, in which the diseases associated with drugs were used for representing drug molecules. Then, we combined disease information with the DrugBank database. Finally, IDF was used to assign more weight and Tanimoto similarity was used to compute disease similarity between a pair of drugs.

MedDRA similarity

The MedDRA defines medical terminology to enable sharing of regulatory information. The terminology is used throughout the regulatory process, from pre-market to post-market, as well as for data entry, consultation, evaluation, and presentation. In the feature matrix, MedDRA is used for side-effect similarities. The side-effect information is gathered from TwoSIDES, which is a source of polypharmacy

```
# Entries for the BLOSUM62 matrix at a scale of ln(2)/2.0.
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  J  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1 -1 -1 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1 -2  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  4 -3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4 -3  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -1 -3 -1 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0 -2  4 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1 -3  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -4 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0 -3  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3  3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4  3 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0 -3  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3  2 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3  0 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -3 -1 -1 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0 -2  0 -1 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1 -1 -1 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -2 -2 -1 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -1 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3  2 -2 -1 -4
B -2 -1  4  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4 -3  0 -1 -4
J -1 -2 -3 -3 -1 -2 -3 -4 -3  3  3 -3  2  0 -3 -2 -1 -2 -1  2 -3  3 -3 -1 -4
Z -1  0  0  1 -3  4  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -2 -2 -2  0 -3  4 -1 -4
X -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Fig. 2. BLOSUM62 – replacement of amino acids

ADRs for combinations of drugs gathered from FDA Adverse Event Reporting System (FAERS) and the DrugBank. We have used IDF weight for the terms and computed Tanimoto coefficient for similarities between pairs of drugs.

Enzyme similarity

Enzyme similarity was computed using the approach applied to target similarity. Amino acid sequences of the enzyme protein were obtained using the UniProt database, and the Smith–Waterman sequence alignment score between drug target genes was computed as suggested by Bleakley and Yamanishi.[23] The geometric mean of the scores was calculated to normalize the score obtained from aligning each sequence.

PPI similarity

Proteins are in charge of all biological systems in a cell, and while many proteins operate on their own, the great majority of them interact with one another to ensure appropriate biological activity. The protein–protein interaction (PPI) network was created, and the closest distance between proteins was computed. Following the suggestions of Rohani and Eslahchi,[7] the distances were converted to similarity measurements using the following equation (Equation 6):

$$S(p1, p2) = A \times e^{-D(p1, p2)} \quad (6)$$

where $S(p_1,p_2)$ is the computed similarity value between 2 proteins, $D(p_1,p_2)$ is the shortest path between these proteins in the PPI network, and A was chosen to be $0.9 \times \exp(1)$. Figure 3 presents one-step interactions between proteins for illustration purposes.

Data analysis

This paper was intended to investigate major cardiovascular drugs and their interactions with other drugs. First, we took few well known drugs and their interaction to other available drugs. Then we added new drug and its interactions with other available. Finally, in the current study, we used 22 drugs in drug 1 column. In the current study, we used 22 drugs as drug 1 and 841 drugs as drug 2, totaling 18,249 data points after removing duplicates. Sample data is shown in Table 1. During the analysis, we used R v. 4.0.2 software (R Foundation for Statistical Computing, Vienna, Austria).

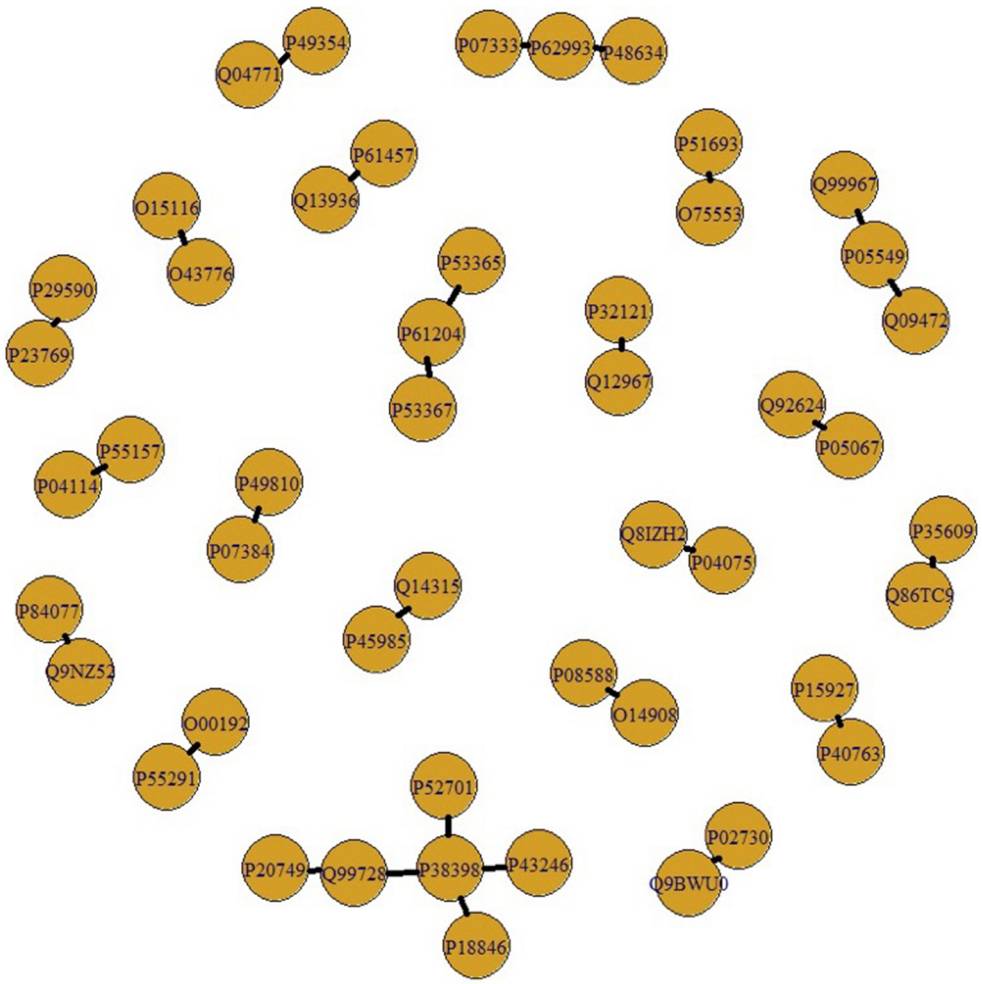

Fig. 3. One-step protein–protein interactions

Checking for multicollinearity is a vital step of data analysis. Multicollinearity indicates which independent variables are not independent of each other, and whether the high correlation between independent variables might cause problems with not only fitting but also interpreting the results. Therefore, in Fig. 4, the correlation matrix between 8 feature matrices can be observed. The highest correlation is 0.44 between Pathway similarity and Disease similarity. This is followed by a correlation 0.38 between Enzyme similarity and PPI similarity.

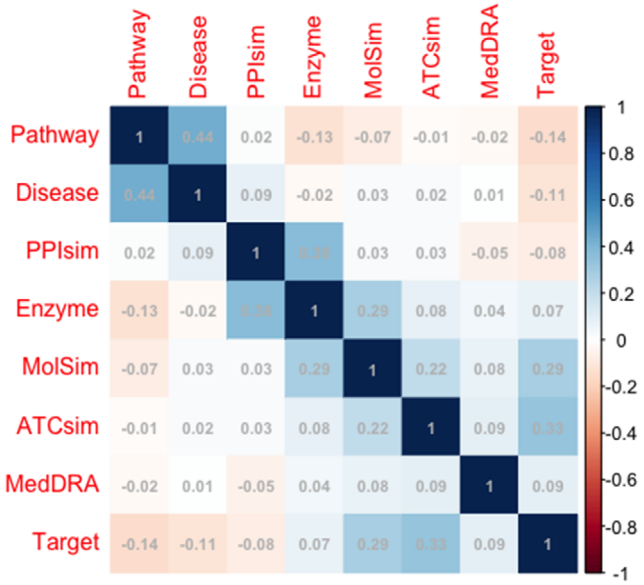

Fig. 4. Correlation matrix between explanatory variables

We checked the assumptions using LR and found that there were no highly influential outliers in the data. Cook's distance was estimated to verify the finding. Linear relationships between each explanatory variable and the logit of the response variables were checked using the Box–Tidwell method. We have fitted a LR model, which contains both x and x*log(x) for all of our explanatory variables (x). We then added those interactions into a model and fitted it to LR. We found out that ATC, target, enzyme, and disease do not satisfy linearity assumptions. Their p-values were <2e-16, <2e-16, <2e-16, and 0.01747, respectively. Later, we created new variables using the square of ATC, target, enzyme, and disease. However, ATC was not significant. In the end, we used the square of the target, enzyme and disease variables, and the natural log of ATC. We checked all LR assumptions for new variables and did not notice any problems. Since we have a large dataset, overfitting was not our primary issue. The goodness-of-fit of the LR model was evaluated by using Nagelkerke's R^2 from the fmsb R package and obtained a value of 0.2199, showing a moderate relationship.

Table 3. Model summary of logistic regression

Parameter	df	Deviance	Resid. Df	Resid. Dev	p-value (>χ)
NULL	–	–	14591	19336	–
Pathway	1	25.88	14590	19311	0.0046
MedDRA	1	473.24	14589	18837	0.0000
MolSim	1	1200.67	14588	17637	0.0000
ATCsim_In	1	251.25	14587	17385	0.0000
Target2	1	323.50	14586	17062	0.0000
Enzyme2	1	275.20	14585	16787	0.0000
PPIsim	1	10.62	14584	16776	0.0022
Disease2	1	10.63	14583	16766	0.0012

df – degrees of freedom; Resid. Df – residual degree of freedom; Resid. Dev – residual deviance.

On the other hand, the model summary presented in Table 3 shows that all covariates are statistically significant.

Classification problems, especially with binary outcomes, were labeled as positive or negative. The decision was made by using a confusion matrix or contingency table, which consist of 4 categories: 1) true positive (TP) occurs when the outcome is successfully classified as positive; 2) true negative (TN) occurs when the outcome is successfully classified as negative; 3) false positive (FP) refers to an outcome as positive when the truth is negative; and 4) false negative (FN) refers to an outcome as negative when the truth is positive. Sensitivity refers to the ability of a test to correctly identify events related to a disease. Specificity, on the other hand, refers to the ability of a test to reliably detect events that occur in the absence of disease. Equations 7–11 has been used to compute these values in Table 2.

$$\text{TPR} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \tag{10}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

A confusion matrix can be obtained after fitting any machine learning algorithm and obtaining predictions on the test set. An accuracy of algorithm can be checked by computing some measures such as false positive rate (FPR) to indicate the fraction of negative groups which are incorrectly classified as positive. True positive rate (TPR) indicates the fraction of positive groups that are successfully classified as positive. Precision indicates the measure of correctly spotted positive events out of the predicted positive events. Recall indicates the measure of correctly spotted positive events out of all the actual positive events. Accuracy indicates the measure of all the correctly spotted events. When the ratio of the positive and negative events is close to each other, accuracy is a good indicator to consider. Although it is not always the case for equally numbered groups, in real datasets, mainly positive and negative events are not equal, which indicates imbalanced data. Therefore, an F1 score would be useful in the case of unequal groups and when FPs and FNs are being considered, since mislabeling interactions as non-interactions might cause patient's death.

The dataset is split into branches (80% training and 20% testing). Three machine learning algorithms were used to fit the data in R software using the macOS Ventura 13.3.1 operating system (Apple Inc., Cupertino, USA). Results are shown in Table 2. We used the xgboost package to fit XGBoost, Keras and Tensorflow packages for fitting the NN and, finally, the stats package was used for fitting the LR. A 10-fold cross-validation was carried out independently for each algorithm to validate our models.

Conclusions

We have used 3 popular machine learning algorithms, LR, XGBoost and NN, for solving a classification problem. Each method has its strengths and weaknesses. Logistic regression is a simple algorithm that is easy to implement, interpret and explain. On the other hand, LR has limitations in its ability to capture complex, non-linear relationships between the dependent and independent variables. The XGBoost and NN are commonly preferred blackbox methods. They have high accuracy and power to handle larger datasets but time-consuming hyperparameter tuning and difficult interpretations.

Following the results presented in Table 2, we could say that using XGBoost would be a good machine learning algorithm in this case. We achieved the highest accuracy among all algorithms using XGBoost (78%), which is not surprising because, as we pointed out before, XGBoost

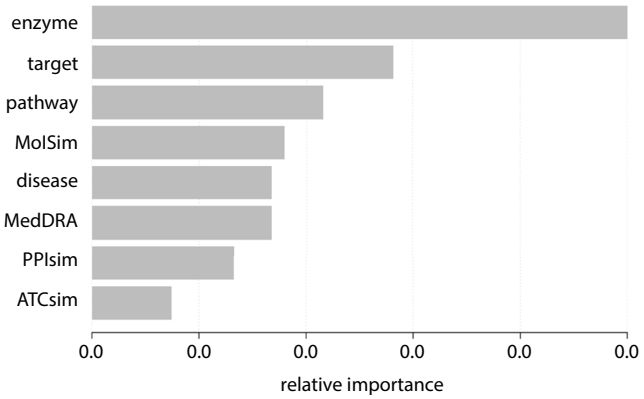

Fig. 5. Feature importance from eXtreme Gradient Boosting (XGBoost)

uses gradient boosting inside the algorithm. Therefore, it is known to retrieve importance scores for each feature. In general, significance assigns a score to each feature that reflects how useful or important it was in the development of the model's enhanced decision trees. The higher the relative relevance of a characteristic, the more often it is used to make critical judgments with decision trees. After the training, xgb.importance was used to identify which features have higher importance.[25] Feature selection is a widely used approach in machine learning.[26,27] Enzyme similarity is the most important feature in the model, followed by target similarity (Fig. 5). Finally, ATC similarity is the least important feature. Thus, even if some papers are only focused on molecular similarity, enzyme and target protein similarities are the most vital features when it comes to identifying DDIs.

Kappa is a statistical measure that evaluates the degree of agreement between predicted and actual values. It ranges from –1 to 1, with 1 indicating perfect agreement, 0 indicating random agreement and –1 indicating complete disagreement. Based on the kappa values reported in Table 2, it is apparent that the XGBoost model exhibits the highest agreement between predicted and actual values, with a kappa value of 0.5508. This finding suggests that the XGBoost model is capable of predicting outcomes more accurately than the other 2 models. In contrast, the LR model reached the lowest kappa value of 0.2641. The NN model, with a kappa value of 0.4154, falls in between the values obtained for the XGBoost and LR models. These results suggest that the NN model is more accurate than the LR model but not as accurate as the XGBoost model in predicting outcomes.

Creating a new feature matrix is computationally requiring more powerful devices; therefore, using excessive feature matrices to identify DDIs in methods would would be interesting to investigate if the purpose is solely for feature selection in DDIs. However, we believe that after feature selection, the model would end up with 8 or 9 explanatory variables, which would be similar to the ones used in this paper. Additionally, we compared 3 machine learning algorithms, but every other classification method available in R programming could be used. Finally, we used 22 drugs as drug 1 and 841 drugs as drug 2 (see Table 1), but it could be expanded to larger numbers with more powerful devices and extra time, as mentioned earlier. Our method creates feature matrices from raw data and is still faster than many other approaches. The method would be vital for scientists in drug development since this is a non-clinical and accurate approach. It would also lower Research and Development (R&D) expenses of pharmaceutical companies. A R-shiny app or another automation system can be created to obtain probability of DDI interaction for chosen drugs. This would help doctors when deciding which drugs to prescribe for a given patient.

Supplementary data

The Supplementary file has 22 folders for each drugs in a drug 1 column in Table 1. The R codes which is used to fit all 3 machine learning algorithms can be found at https://github.com/iDemirsoy/Understanding-DDI-.

ORCID iDs

Idris Demirsoy https://orcid.org/0000-0002-3321-4748
Adnan Karaibrahimoglu https://orcid.org/0000-0002-8277-0281

References

1. Percha B, Altman RB. Informatics confronts drug–drug interactions. *Trends Pharmacol Sci*. 2013;34(3):178–184. doi:10.1016/j.tips.2013.01.006
2. Zhang P, Wang F, Hu J, Sorrentino R. Label propagation prediction of drug–drug interactions based on clinical side effects. *Sci Rep*. 2015;5:12339. doi:10.1038/srep12339
3. Vilar S, Friedman C, Hripcsak G. Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media. *Brief Bioinform*. 2018;19(5):863–877. doi:10.1093/bib/bbx010
4. Vilar S, Uriarte E, Santana L, et al. Similarity-based modeling in large-scale prediction of drug–drug interactions. *Nat Protoc*. 2014;9(9):2147–2163. doi:10.1038/nprot.2014.151
5. Fokoue A, Sadoghi M, Hassanzadeh O, Zhang P. Predicting drug–drug interactions through large-scale similarity-based link prediction. In: Sack H, Blomqvist E, d'Aquin M, Ghidini C, Ponzetto SP, Lange C, eds. *The Semantic Web. Latest Advances and New Domains*. Cham, Switzerland: Springer International Publishing; 2016:774–789. doi:10.1007/978-3-319-34129-3_47
6. Wu HY, Chiang CW, Li L. Text mining for drug–drug interaction. In: *Biomedical Literature Mining*. New York, USA: Springer New York; 2014:47–75. doi:10.1007/978-1-4939-0709-0_4
7. Rohani N, Eslahchi C. Drug–drug interaction predicting by neural network using integrated similarity. *Sci Rep*. 2019;9(1):13645. doi:10.1038/s41598-019-50121-3
8. Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, Friedman C. Drug–drug interaction through molecular structure similarity analysis. *J Am Med Inform Assoc*. 2012;19(6):1066–1074. doi:10.1136/amiajnl-2012-000935
9. Roos J. Cardiac effects of antidepressant drugs: A comparison of the tricyclic antidepressants and fluvoxamine. *Br J Clin Pharmacol*. 1983;15(Suppl 3):439S–445S. doi:10.1111/j.1365-2125.1983.tb02135.x
10. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform*. 2015;7:20. doi:10.1186/s13321-015-0069-3
11. Papineni K. Why inverse document frequency? In: Second Meeting of the North American Chapter of the Association for Computational *Linguistics*; 2001. https://aclanthology.org/N01-1004.pdf. Accessed August 15, 2020.

12. Church K, Gale W. Inverse Document Frequency (IDF): A measure of deviations from Poisson. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, eds. *Natural Language Processing Using Very Large Corpora*. Dordrecht, the Netherlands: Springer Netherlands; 1999:283–295. doi:10.1007/978-94-017-2390-9_18
13. Chen T, He T, Benesty M. Xgboost: Extreme gradient boosting. Melbourne, Australia: University of Melbourne; 2015. https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf. Accessed August 11, 2020.
14. Demirsoy I. Estimating the Intensity of Point Processes on Linear Networks [doctoral thesis]. Tallahasee, USA: Florida State University; 2020. https://diginole.lib.fsu.edu/islandora/object/fsu:781648. Accessed August 15, 2020.
15. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA: ACM; 2016: 785–794. doi:10.1145/2939672.2939785
16. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. *Heliyon*. 2018;4(11):e00938. doi:10.1016/j.heliyon.2018.e00938
17. Paul S, Singh A. *Hands-On Python Deep Learning for the Web*. Birmingham, UK: Packt Publishing; 2020. ISBN:978-1-78995-608-5.
18. Agarap AF. Deep learning using rectified linear units (ReLU). *arXiv*. 2018. doi:10.48550/ARXIV.1803.08375
19. Wanto A, Windarto AP, Hartama D, Parlina I. Use of binary sigmoid function and linear identity in artificial neural networks for forecasting population density. *IJISTECH*. 2017;1(1):43. doi:10.30645/ijistech.v1i1.6
20. Peng CYJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. *J Educ Res*. 2002;96(1):3–14. doi:10.1080/00220670209598786
21. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug–target interactions: A brief review. *Brief Bioinform*. 2014;15(5):734–747. doi:10.1093/bib/bbt056
22. Landrum G. Getting started with the RDKit in Python: The RDKit 2023.03.1 documentation. 2018. https://www.rdkit.org/docs/GettingStartedInPython.html. Accessed August 11, 2020.
23. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*. 2009;25(18): 2397–2403. doi:10.1093/bioinformatics/btp433
24. Noh K, Yoo S, Lee D. A systematic approach to identify therapeutic effects of natural products based on human metabolite information. *BMC Bioinformatics*. 2018;19(Suppl 8):205. doi:10.1186/s12859-018-2196-0
25. Chen T, He T, Benesty M, Khotilovich V. Package 'Xgboost'. Toronto, Canada: University of Toronto; 2022. https://cran.utstat.utoronto.ca/web/packages/xgboost/xgboost.pdf. Accessed August 15, 2020.
26. Chen C, Zhang Q, Yu B, et al. Improving protein–protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Comput Biol Med*. 2020;123:103899. doi:10.1016/j.compbiomed.2020.103899
27. Wang Y, Ni XS. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv*. 2019. doi:10.48550/ARXIV.1901.08433