Predicting Adverse Drug Events Using Network Metrics

Natalie Wang, Kritika Sharma, Elahe Mohammadi Siahroodi

Abstract

One of the most significant problems in public health is ensuring the safety of drugs by understanding and being able to predict their adverse effects or ADEs (Adverse Drug Events). Motivated by increases in hospitalizations and deaths due to ADEs and the poor performance of traditional methods and with the growth of drug and biological data repositories, researchers have turned to machine learning and data science methods to aid in the prediction of ADEs. In this research, we utilize network metrics such as similarity and centrality from a novel drug-drug-ADE network along with machine learning techniques to predict associations between drugs and heart-related ADEs. An extensive collection of annotated MEDLINE biomedical articles was used to construct a drug-ADE network, and the network was further equipped with information about drugs' target proteins. We created a novel drug-drug-ADE network and trained accurate machine learning classifiers to predict drug-ADE links based on network metrics. Our top model was a bagged decision tree that identified drug-ADE associations with an overall accuracy of 90.6%.

1 Introduction

Our project investigates drug-protein-ADE (Adverse Drug Event) relations to identify similar drugs based on the proteins they target and how they are associated with the drugs' reported ADEs. We aim to explore newer drug compared to existing drugs with well-known ADEs by using target protein commonalities to predict likely new drug-ADE relationships. We hope our method can aid ADE detection in the clinical trial stage or before, thus averting severe and often fatal adverse effects that are experienced by users when drugs are approved prematurely.

Accurately identifying adverse drug events early is an increasing concern in the medical industry. Medical errors have been the cause behind the death and injury of over one million patients in the US alone. ADEs contribute about a fifth of that [1]. Even though there are several surveillance techniques in practice to monitor ADEs, studies indicate that the best review technique is chart monitoring [1], which is time and resource expensive. Hence, several automated learning techniques have been employed to tackle this issue. The most impactful research in this field has been done by Cami et al. [2], where pharmaco-safety networks (PPNs) have been employed. In PPNs, known drug-ADE relationships on specific drugs are used to predict likely unknown ADEs. The crux of this predictive approach relies on leveraging existing, contextual drug safety information, potentially identifying certain ADEs very early.

We'll investigate drug-target-ADE relations to identify target proteins associated with the reported ADEs. One popular approach [3] uses ADE vectors and computes a cosine correlation coefficient to different drugs. Another popular approach is the chemogenomic approach [4] which computes drug-drug similarity by evaluating chemical structural similarities between drugs. In our project, we have adopted a mixed network science and machine learning approach to the problem where drug-drug similarity is calculated by comparing the target proteins and network metrics are used to train a classifier for drug-ADE links.

2 Background

2.1 Computational Medicine

In recent years, with the explosive growth of biomedical data and the rapid development of medical and

computer technology, the field of digital health was ushered in the era of big data [5]. In this context, computational medicine began to appear as a new subject. Based on big biomedical data and computer technology, computational medicine is an interdisciplinary subject combining medicine, computer science, biology, mathematics, etc [6]. One of the prevailing direction in this area is computational pharmacology; three specific aims within the realm of computational pharmacology are prediction of drug-target interactions, prediction or explanation of potential side effects or adverse drug reactions, and methods for drug re-purposing, i.e. finding new uses for existing drugs [7].

2.1.1 ADEs

The drugs in adverse *drug* events are defined by the FDA as "A substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease". Therefore, in this study, we are looking at medications prescribed by doctors and not drugs such as cocaine or heroin. Adverse drug events or ADEs are any unexpected harm caused by the normal use of medication at the normal dosage. It therefore does not cover those injuries caused by inappropriate or off-label usages of a medication [8]. ADEs cause hospitalizations and mortality in substantial numbers, and their incidence is on the rise [9][10][11], which encourages researchers to develop new detection and prediction methods.

ADE studies can be classified into 3 distinct categories, namely, detection, prediction, and understanding [12]. Detection studies are the largest group of ADE research works focused on finding new and undetected ADE signals (ie, associations, not necessarily causal) between the existent drugs (already in the market) and adverse events [13]. In ADE prediction studies, instead of detecting signals for the existent drugs using collected data from their past usage experiences, the focus is on creating signals for the new drugs before they cause any ADEs to patients. The strategy in this group of studies is mainly to find similarities between the existent and the new drugs and thereby to predict ADEs for the new drugs given the already known relationships between their similar existent drugs with the corresponding ADEs [14][15]. We will be working in this second category of studies. The last group of ADE studies in our taxonomy are those focusing on verifying ADE signals and understanding the mechanism through which the drug causes the ADE [16].

2.2 Network Analysis

2.2.1 Network Metrics

If we know a network's structure, we can calculate a variety of useful quantities or measures that capture particular features of the network topology [17].

The centrality of a node in a graph defines how important a node in the graph is. There are a wide variety of mathematical measures of vertex centrality that focus on different concepts and definitions of what it means to be central in a network.

Consider network $N = (V, E, w)$ with adjacency matrix A, where $A_{ij} = 1$ if there is an edge between nodes i and j in V; and $A_{ij} = 0$ if there does not exist an edge between nodes v_i and v_j.:

Definition 1 (Degree Centrality) *Degree centrality is the most straightforward network centrality measure. It only considers the degree of a node, which is the number of nodes that a given node is connected to [18].*

$$c_d(v_i) = k_i \quad \forall v_i \in V \tag{1}$$

Degree centrality measures the extent of influence that a node has on the network. The more neighbors a node has, the more critical it is. Although the concept of degree centrality is straightforward, sometimes we need to find some central nodes with more complicated structures.

Definition 2 (Closeness Centrality) *The closeness centrality of a node measures the centrality of a node based on how close it is to other nodes in the network. The smaller the total distance of a node to other nodes, the higher its closeness is. The distance between two nodes is defined as the shortest path length between them. We calculate the closeness centrality measure for a node by inverting the sum of the distances from it to other nodes in the network [18].*

Similarity metrics compare how similar or equivalent two nodes (or more complex structures within the graph) are. It is a tool that can help predict the association between two nodes.

Definition 3 (Jaccard Similarity) *Jaccard coefficient of nodes $v_i, v_j \in V$ is defined as*

$$J(v_i, v_j) = \frac{|\Gamma(v_i) \cap \Gamma(v_j)|}{|\Gamma(v_i) \cup \Gamma(v_j)|} \tag{2}$$

Other network metrics, the Adamic-Adar Coefficient and the Preferential Attachment Coefficient, which we chose to use in this project, are proposed to be helpful for link prediction.
In 2003, Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar introduced a measure which predicts the association between two nodes based on their similar neighbours [19]:

Definition 4 (Adamic–Adar Coefficient) *For each two nodes $v_i, v_j \in V$ define the Adamic–Adar Coefficient:*

$$A(v_i, v_j) = \sum_{u \in \Gamma(v_i) \cap \Gamma(v_j)} \frac{1}{ln|\Gamma(u)|} \tag{3}$$

where $\Gamma(w)$ is the set of neighbours of $w \in V$.

The definition is based on the concept that common elements with substantial neighborhoods are less significant when predicting a connection between two nodes compared with elements shared between a small number of nodes.

In 1999, Albert L. Barabasi and Reka Albert introduce the preferential attachment score [20] which indicates new vertices added to the graph are attached preferentially to high degree vertices.

Definition 5 (Preferential Attachment) *For each two nodes $v_i, v_j \in V$ define the preferential attachment score:*

$$P(v_i, v_j) = |\Gamma(v_i)||\Gamma(v_j)| \tag{4}$$

where $\Gamma(w)$ is the set of neighbours of $w \in V$ [21].

2.2.2 Machine Learning Techniques for Prediction

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. There are various algorithms for classification; here, we review some of them and apply them to our problem.

Logistic Regression can be viewed as arising from a Bernoulli model. Given a set of predictors, x_n, we wish to determine the probability of a binary out- come Y_n. We define a probability model:

$$\mathbb{P}(Y_n = 1|x_n) := \sigma(wx_n)$$

with corresponding likelihood function:

$$\mathbb{P}[y|x_n, n = 1, ..., N) = \prod_n \sigma(wx_n)^{Y_n}(1-\sigma(wx_n))^{1-Y_n}$$

where the logistic function

$$\sigma(\theta) = \frac{1}{1 + exp(-\theta)}$$

is a continuous increasing function mapping any real valued θ into the interval $(0, 1)$, and thus is suitable for representing the probability of a Bernoulli trial outcome [22].

Naïve Bayes classifies using Bayes' Theorem of probability.

Theorem 1 *For events A and B, if the probability of B is non zero, $\mathbb{P}[B] \neq 0$ then we have:*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

where $\mathbb{P}(.|.)$ is a conditional probability.

Naïve Bayes' classifiers fall under the category of simple probabilistic classifiers based on the concept of Bayes' Theorem having strong independence assumptions among the features. It is particularly suited when the dimensional of the inputs is high [23][24].

K-nearest Neighbours is a non-parametric method used for classification and regression. Given N training vectors, the KNN algorithm identifies the k-nearest neighbors of an unknown feature vector whose class is to be identified.

Support Vector Machines SVM works on the concept of margin calculation. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features we have in our data set). The value of each feature is the value of the corresponding coordinate. It classifies the data into different classes by finding a line (hyperplane) that separates the training data sets into classes. It works by maximizing the distances between the nearest data point (in both classes) and the hyperplane that we can call as margin [25].

Decision Trees is a technique for approximating discrete-valued target function, which represents the learned function in the form of a decision tree [26]. A decision tree classifies instances by sorting them from root to leaf nodes based on feature values. Each node represents some decision (test condition) on the instance's attribute, whereas every branch represents a possible value for that feature. While using a decision tree, the focus is on deciding which attribute is the best classifier at each node level. Statistical measures like information gain, Gini index, Chi-square, and entropy are calculated for each node to calculate the worth of that node [26].

Gradient Boosted Trees is a technique where an ensemble of weak learners is used to improve the performance of a machine-learning model. The weak learners are usually decision trees—combined, their output results in better models [27].

Bagged Decision Trees is a technique in which an ensemble of weak learners trained on randomly selected subsets of the training set is aggregated to create one prediction. Aggregating predictions of different models can reduce variance and overfitting [28].

Random Forest uses a bagging approach to create a bunch of decision trees with a random subset of data. The output of all decision trees in the random forest is combined to make the final decision trees. There are two stages in Random Forest Algorithm, one is to create a random forest, and the other is to predict the random forest classifier created in the first stage [25].

3 Materials and Methods

3.1 Materials

3.1.1 ADE Datasets

We used SIDER, Side Effect Resource, an open dataset that provides information on marketed medicines and their recorded adverse drug reactions. The database has been scraped from public documentation and drug labels. It includes information for each drug such as side effects, the side effect in MedDRA preferred terms, side effect frequency, and indications (why that medicine would be prescribed). MedDRA is the Medical Dictionary for Regulatory Activities. It is an internationally used set of terms for medical conditions, medicines, and devices created to help standardize medical information. The indications for each medicine were scraped from the indications and usage sections of the labels.

3.1.2 Protein Datasets

DrugBank Online is a freely accessible online database that contains information on drugs and drug targets. It contains detailed drug data and drug target information. It currently contains 14,975 drug entries with 5,290 non-redundant protein sequences linked to these entries. We queried DrugBank Online to create our drug-protein dataset.

3.1.3 MEDLINE Datasets

MEDLINE is a database from the National Library of Medicine. It contains more than 12 million bibliographic citations from thousands of biomedical journals. We used PubMed to query the year drug-ADE

associations were first found. We used this information to separate our training and validation datasets.

3.1.4 Identifying Training and Validation Datasets

Using the MEDLINE database, we used drug-ADE pairs found between 2000-2014 as our training set and pairs found between 2015-2022 as our test set. We had about 3600 training pairs and 490 test pairs. We were thus able to attempt to predict ADEs of newer drugs with our ML models and check our predictions with the published drug-ADE pairs discovered after 2015.

3.2 Methods

3.2.1 Network Construction

Our network was a drug-drug-ADE network. We chose to filter our original drug-ADE dataset from SIDER to only the top 10 most common heart-related ADEs. These included: Cardiac Disorder, Myocardial Infarction, Tachycardia, Supraventricaular Tachycardia, Myocardial Ischaemia, Bradycardia, Cardiac Failure, Cardiac Arrest, Ventricular Tachycardia, and Cardiac Failure Congestive. We also chose to only look at proteins that were targeted by at least 5 drugs; this allowed us to filter out individual drugs that were not particularly similar to any other drugs in the dataset. We ended up with 409 drugs, 10 ADEs, and 9,106 links between drugs and ADEs.

Our final network was similar to a bipartite graph with drugs on one side and heart-related ADEs on the other except we also had links between drugs. We added a link between two drugs if they shared at least one target protein, and a link between a drug and an ADE if there existed a mention of that drug causing the particular ADE. It was an unweighted, undirected graph.

From our drug-drug-ADE network, we projected just the drug-drug part of the network to a new graph to examine the drug-drug relations closer. We found the drugs to be clustered in 5 groups based on their protein targets and chose to do the rest of the analysis on the largest group that contained most of the drugs.

3.2.2 Network Metrics

For every possible drug-ADE pair (409 drugs $\times$ 10 heart-related ADEs), we calculated 5 network metrics. We calculated two centrality metrics: degree centrality and closeness centrality. For the degree centrality, we summarize the degree centrality of the drug in the pair and the degree centrality of the ADE in the pair. We did the same for the centrality of closeness. Additionally, we calculated 3 similarity metrics: Jaccard coefficient, Adamic adar coefficient, and preferential attachment coefficient for each drug-ADE pair. We used these based on other literature that proposed that they may be helpful in link prediction [12].

3.2.3 Machine Learning Models

We used our similarity metrics to train machine learning models. Our input X for each classifier was each drug-ADE's five network metrics, and our input y for each classifier was a binary variable, either 1 or 0, indicating whether or not there was a link between that drug-ADE pair. Refer to Figure 5 as an example of the dataset we used for our classifier.

Using our network metrics, we trained eight classifiers: logistic regression, Naive Bayes, KNN, SVM, decision tree, bagged decision tree, boosted decision tree, and a random forest on our training set of drug-ADE pairs found before 2015 and any possible drug-ADE pairs for drugs approved before 2015. We then validated our models by predicting drug-ADE links for drug-ADE links discovered post-2015.

4 Results

In this section, we will illustrate our results in some graphs and tables. First, there are samples of our graphs, shown in Figures 1, 2, and 3. Due to the large size of the network (409 drugs, 10 ADEs, and 9,106 links), it is hard to visualize and comprehend the complete network. Therefore, we created a subsection of the graph from 10 drugs and 8 heart-related ADEs to demonstrate the network (Figure 2).

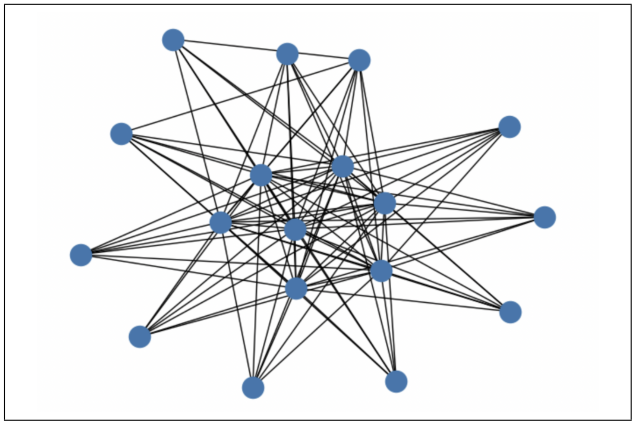

Figure 1: The Drug-Drug Network with 18 drugs (nodes) based on common target Protein

Figure 1 shows the Drug-Drug Network with 18 drugs (nodes). The nodes are drugs and there exists a link between two drugs if they share at least one common protein target based on the DrugBank dataset.

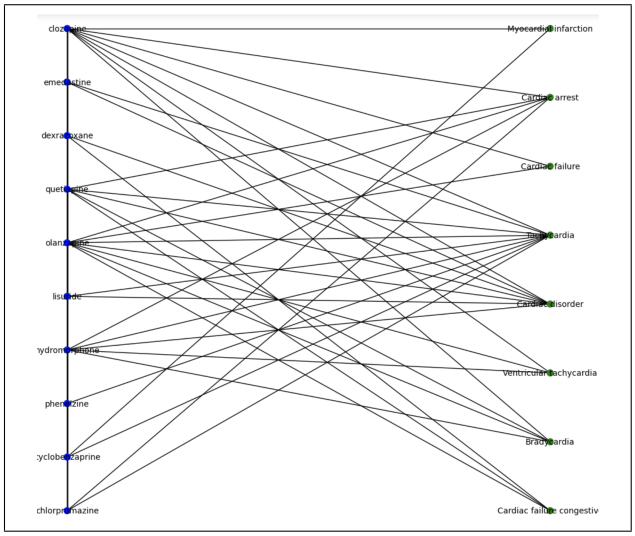

Figure 2: The Drug-Drug-ADE Network with 10 drugs and 10 ADEs (nodes)

Figure 2 shows the Drug-Drug-ADE network. The blue nodes on the left side are drugs and the green nodes on the right side are ADEs. Links between drugs show an association between two drugs if there exists at least one common protein target and links between drugs and ADEs show the drug-ADE associations found from the SIDER dataset.

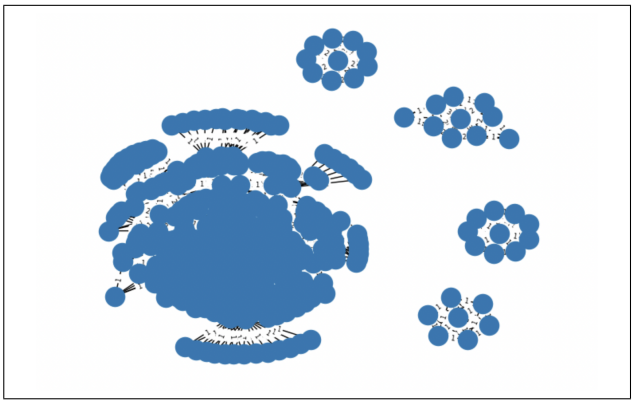

Figure 3: The Drug-Drug network based on common target Protein with five biggest groups

Figure 3 is a schematic picture of drug-drug network construct based on common protein targets. The graph consists of five distinct groups once we filtered out proteins that were targeted by less than five drugs. After looking at this graph, we chose to only consider drugs in the largest group for our classifier to ensure they would all have some similar targets.

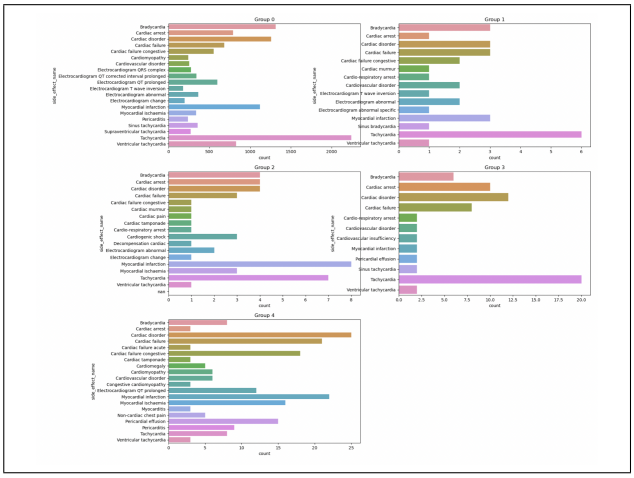

Figure 4: The frequencies of ADEs in each group of drugs in figure 3

Figure 4 shows the frequencies of ADEs in each

group of drugs in figure 3. Some of the ADEs are present in every group like Bradycardia; some are highly prevalent, e.g. Tachycardia. However, some are only seen in a few groups, such as pericarditis which we can see only in Group 0 and Group 4.

	drug	ade	link	jaccard	adamic_adar	pref_attach	deg	close	year	first_mention
0	bupropion	Cardiac disorder	1	0.109649	5.795098	11310	0.483748	1.095237	2012.0	2000.0
1	bupropion	Myocardial infarction	1	0.140541	5.884129	8874	0.403442	1.058802	2000.0	2000.0
2	bupropion	Tachycardia	1	0.160772	11.707961	17574	0.690249	1.190730	2002.0	2000.0
3	bupropion	Supraventricular tachycardia	1	0.098039	2.359099	3132	0.214149	0.969717	2008.0	2000.0
4	bupropion	Myocardial ischaemia	0	0.095238	2.444771	3306	0.219885	0.981269	0.0	2000.0

Figure 5: The slice of data with calculated network metrics for classification

In figure 5, you can see a slice of data used for the classification model. The first two columns indicate drugs and ADEs. A drug-ADE pair is classified as zero or one in the third column depending if there was a published mention of that specific drug causing that ADE. The other columns show the network metrics calculated; these were the classification features.

	model	train_score	test_score	train_time
5	bagged decision tree	0.9856233333333330	0.9061224489795920	0.08675880299779240
6	boosted decision tree	0.8178455555555560	0.9040816326530610	0.4473036529998350
7	random forest	0.99979	0.9040816326530610	0.508588590000727
1	naive bayes	0.6717344444444440	0.8551020408163270	0.002927118999650700
3	SVM	0.6634011111111110	0.8551020408163270	0.5559452199995580
0	logistic regression	0.6722900000000000	0.8489795918367350	0.02799344799859680
4	decision_tree	0.99979	0.8306122448979590	0.012862640000094000
2	knn	0.76979	0.7591836734693880	0.0027567020006245

Figure 6: The summary of the performance of the different classification algorithms.

According to their training and test accuracy and training run time, Figure 6 shows the performance of our eight different classification algorithms. Overall, the *decision tree* algorithms have better results than the others; the best being the bagged decision tree.

5 Discussion

We noted that by considering drug-target interactions with additional network variables, we were able to predict drug-ADE links post-2015 with a maximum accuracy of 90.6%. Our results are more accurate compared to previous literature, [12]. One possible reason may be that we focused on only the top ten most common heart-related ADEs, whereas the previous work looked at the top eight most common and high-risk ADEs. As a result, there may be a correlation between our heart-related ADEs, which can lead to more accurate prediction results. Furthermore, using more recent and accurate datasets helped us improve our model. Some prediction errors can be attributed to more complex drug-drug relations; for example, two or more drugs may still interact while targeting different proteins. More intensive and complex algorithms, such as deep learning methods, may be better at modeling drug-drug inter-effects, thus improving drug-ADE prediction.

Our main insights from this research were more global than local. We created a novel drug-drug-ADE network and trained accurate machine learning classifiers to predict drug-ADE links based on network metrics. Our network and these models can be used in the pharma-industry to predict heart-related ADEs on newly-approved drugs.

6 Conclusion

In our project, we predict ADEs by constructing an unweighted drug-drug-ADE networks using biomedical citations and drug target proteins information. We employed network metrics in addition to several classical machine learning techniques to predict associations. Our results suggest that using centrality-based metrics (degree centrality and closeness centrality), similarity-based metrics (Jaccard coefficient), and link prediction metrics (Adamic-Adar coefficient and preferential attachment coefficient) together can be beneficial for such classifying tasks. Additionally, all the information used to train our prediction models were historical drug-ADE associations and drug-target proteins.

6.1 Future Work

One way to improve our results may be to increase the size of our training dataset. In the future, we would like to extend our dataset by combining information from multiple public databases such as AERS,

SIDER, JAPIC, etc., as well as scraping public forums like Facebook and Twitter. This may help improve the accuracy of ADE prediction. Additionally, considerations around ethnicity, gender, medical history, economic background, etc., may be crucial factors in improving the accuracy of drug-ADE predictions. Considering higher-order drug-drug and drug-ADE interactions may help improve our models. We would also like to extend our network by incorporating genomic data and trying different subsets of ADEs such as respiratiory-related ADEs and psychological ADEs.

References

[1] Ashish K Jha, Gilad J Kuperman, Jonathan M Teich, Lucian Leape, Brian Shea, Eve Rittenberg, Elisabeth Burdick, Diane Lew Seger, Martha Vander Vliet, and David W Bates. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association*, 5(3):305–314, 1998.

[2] Aurel Cami, Alana Arnold, Shannon Manzi, and Ben Reis. Predicting adverse drug events using pharmacological network models. *Science translational medicine*, 3(114):114ra127–114ra127, 2011.

[3] Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto, and Yoshihiro Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, 2012.

[4] Yoshihiro Yamanishi. Chemogenomic approaches to infer drug–target interaction networks. *Data Mining for Systems Biology*, pages 97–113, 2013.

[5] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[6] Sijie Yang, Fei Zhu, Xinghong Ling, Quan Liu, and Peiyao Zhao. Intelligent health care: Applications of deep learning in computational medicine. *Frontiers in Genetics*, page 444, 2021.

[7] Rachel A Hodos, Brian A Kidd, Shameer Khader, Ben P Readhead, and Joel T Dudley. Computational approaches to drug repurposing and pharmacology. *Wiley interdisciplinary reviews. Systems biology and medicine*, 8(3):186, 2016.

[8] Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39, 2015.

[9] Munir Pirmohamed, Sally James, Shaun Meakin, Chris Green, Andrew K Scott, Thomas J Walley, Keith Farrar, B Kevin Park, and Alasdair M Breckenridge. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj*, 329(7456):15–19, 2004.

[10] Greene Shepherd, Philip Mohorn, Kristina Yacoub, and Dianne Williams May. Adverse drug reaction deaths reported in united states vital statistics, 1999-2006. *Annals of Pharmacotherapy*, 46(2):169–175, 2012.

[11] Tai-Yin Wu, Min-Hua Jen, Alex Bottle, Mariam Molokhia, Paul Aylin, Derek Bell, and Azeem Majeed. Ten-year trends in hospital admissions for adverse drug reactions in england 1999–2009. *Journal of the Royal Society of Medicine*, 103(6):239–250, 2010.

[12] Behrooz Davazdahemami and Dursun Delen. A chronological pharmacovigilance network analytics approach for predicting adverse drug events. *Journal of the American Medical Informatics Association*, 25(10):1311–1321, 2018.

[13] Ruichu Cai, Mei Liu, Yong Hu, Brittany L Melton, Michael E Matheny, Hua Xu, Lian Duan, and Lemuel R Waitman. Identification

of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. *Artificial intelligence in medicine*, 76:7–15, 2017.

[14] Liang-Chin Huang, Xiaogang Wu, and Jake Y Chen. Predicting adverse side effects of drugs. *BMC genomics*, 12(5):1–10, 2011.

[15] Liang-Chin Huang, Xiaogang Wu, and Jake Y Chen. Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics*, 13(2):313–324, 2013.

[16] Mirjam N Trame, Konstantinos Biliouris, Lawrence J Lesko, and Jerome T Mettetal. Systems pharmacology to predict drug safety in drug development. *European Journal of Pharmaceutical Sciences*, 94:93–95, 2016.

[17] Mark Newman. *Networks*. Oxford university press, 2018.

[18] Linton C Freeman, Douglas Roeder, and Robert R Mulholland. Centrality in social networks: Ii. experimental results. *Social networks*, 2(2):119–141, 1979.

[19] Lada A Adamic, Orkut Buyukkokten, and Eytan Adar. A social network caught in the web. 2003.

[20] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[21] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

[22] Andrew I Schein and Lyle H Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

[23] N Sandhya and KR Charanjeet. A review on machine learning techniques. *International Journal on Recent and Innovation Trends in Computing and Communication*, 4(3):451–458, 2016.

[24] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[25] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series*, volume 1142, page 012012. IOP Publishing, 2018.

[26] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[27] Gradient boosted decision tree classifier. https://neptune.ai/blog/gradient-boosted-decision-trees-guide. Accessed: 2022-12-09.

[28] Byron P Roe, Hai-Jun Yang, and Ji Zhu. Boosted decision trees, a powerful event classifier. In *Statistical problems in particle physics, astrophysics and cosmology*, pages 139–142. World Scientific, 2006.