A Predictive Model for Drug-Drug Interaction Using a Similarity Measure

Abirami Ariyur Mahadevan Anagha Vishnuvajjala Naman Dosi Shrisha Rao

***Abstract*—Drug-drug interaction causes potential impact on patients when a second drug is administered during the duration of action of the first. It may result in the delay or decrease in the absorption of rate of drugs or enhance their absorption. This also in turn may affect the action of drugs or induce adverse effects on patients. There exists a need to study the drug-drug interactions, and their potential effects on the human system, including for drugs not yet approved. This paper proposes using eight features (substructure, targets, transporters, enzymes, pathways, indications, side-effect and off-side-effect, obtained from five different databases - PubChem, Drugbank, KEGG, SIDER, Offsides) and a similarity-based ensemble prediction model to identify the potential drug-drug interactions. The proposed ensemble model uses the Jaccard's coefficient method for identifying similarity measures between drugs. This similarity indices are given to a neighbor recommender method and random walk method for the base prediction of drug-drug interaction. This predictive model is improved by an ensemble model by using a genetic algorithm for weight calculation, and logistic regression for classification. The empirical results show that the ensemble model yields >90% accuracy while predicting the drug-drug interactions.**

***Index Terms*—Drug-drug interactions, machine learning, similarity measures, Jaccard's coefficient, ensemble model, random walk method, neighbor recommender method**

I. Introduction

Drugs are constantly being sought to fight diseases. However, drugs have serious downsides, in terms of side-effects as well as interactions with other drugs. Drug-drug interactions (DDIs) are very common. Some are beneficial to patients, like antidotes administered after overdoses, and drugs used to combat undesirable side-effects of others used in treatment of serious diseases. However, some are potentially harmful and have to be identified at an early stage. Some interactions may have low risk and may be of little clinical significance. It can take years to clinically check DDIs for every pair of drugs. Sometimes DDIs may not get detected in clinical trials. Moreover, it may take many years to check DDIs of all known drugs with a newly discovered one before it is introduced to the market. Hence, there exists the need for efficient DDI-checking with fewer time-consuming, expensive, and risky clinical trials.

Drug-drug interaction occurs when two drugs which are co-administered interact and cause an adverse reaction or unexpected side effects. It can be caused through prescribed medicines, overdose and/or by prolonged use of medicines. Between 2009 and 2012, 38.1% of U.S. adults aged 18–44 used three or more prescription drugs during a 30-day time period [1]. The percentage of drug usage increases substantially with age, becoming 67.2% for ages 45–64, and 89.8% for age 65 years or older respectively. The number of incidents of adverse drug reactions increases exponentially, if a patient takes four or more drugs [2]. However, identifying all possible interactions between all drugs is computationally intractable.

978-1-7281-1462-0/19/$31.00 ©2019 IEEE

DDI detection and remediation requires domain knowledge and the competence to act without undue mental stress to patients and caregivers. Investigations to clinically observe drug interactions are undertaken before marketing, and may assist pharmaceutical companies as well as physicians in gaining confidence about drugs.

Some labor-intensive techniques like *in-silico* methods, *in-vitro* methods, *in-vivo* experiments, and clinical trials may identify DDIs, but they are time-consuming [3]. Statistical methods and machine learning methods were developed to detect the adverse reactions of drugs and drug-drug interactions by analyzing health reports and records. Researchers have also used drug data from literature and health reports and created public databases in order to facilitate the development of classification and prediction methods [3].

Testing all drugs under all possible conditions is impractical and unethical also, hence machine learning is sought to be used. Among machine learning methods that can be used to predict DDIs, there can be two approaches: similarity-based methods, and classification-based methods. In either case, a model is to be created to analyze how drugs interact with other drugs, and used to predict how a new drug would interact with a known one. Similarity-based models assume that similar drugs interact leading to DDIs. Classification-based models consider DDI prediction as a binary classification task in which they use two kinds of data; drug pairs that cause DDIs and drug pairs that do not cause DDIs. In the binary classification, positive labels are given to known interactions between the two drugs; the interactions between other pairs of drugs to be detected using the prediction model.

In this paper we choose similarity-based DDIs because many times the consequences (side effects) of two drugs add up and lead to a DDI. Sometimes similar drugs work in a similar way leading to a DDI because the body cannot sustain both the drugs at the same time.

The Anatomical Therapeutic Chemical classification system (ATC) was used in order to characterize the adverse drug-drug interactions and predict their potential interactions [4].

Fingerprint-based drug-drug interaction model using molecular structure similarity information yields results like 68% sensitivity and 90% specificity [5], [6]. Different drug-drug similarity measures like Chemical-based, Ligand-based, Side-effect based, Annotation-based, and Sequence-based are used to identify the drug-drug interactions which result in 93% sensitivity and specificity values [7]. Pharma co-interaction neural models exist for the prediction of unknown drug-drug interactions by using the logistic regression and generalized linear mixed model with 69% accuracy [8]. A Bayesian probabilistic model has been developed with an accuracy of 82% and recall value 62% for the prediction of pharmacodynamic (PD) DDIs [9]. That model also helped for better understanding on the potential molecular mechanisms or physiological effects underlying DDIs, Heterogeneous network-assisted inference (HNAI) framework was developed using machine learning algorithms like KNN, NB, SVM, Logistic Regression with the accuracy in the range of 60-67% for the prediction of drug-drug interactions. Among these, SVM performs better (with 66.67% accuracy) than other algorithms [10].

In related work, a predictive model has been developed for pharmacodynamic drug-drug interactions using shortest-path-length average method (SPA) for comparison, and random walk with restart algorithm (RWR) technique which yields 80–86% accuracy [11]. A label propagation (LP) algorithm and nearest-neighbor strategies have also been used, and it is found that LP performs better, with the accuracy of 86% for the prediction of drug-drug interactions [12]. A probability ensemble approach (PEA) was developed for analysis of both the efficacy and adverse effects of drug combinations through a Bayesian network model integrated with similarity measures, with the accuracy of 90% [13]. A Web-based demonstration was tried for drug-drug interactions, using different similarity measures (like CPI profile-based, action-based, pathway-based, and ATC-based) and logistic regression model for the 1014 features of dataset collected from DrugBank, and BioGRID [14].

A computational framework for the prediction of of DDIs based on inner-product-based similarity measures was developed and identified 250,000 potential interactions out of 2,394,766 drug pairs [15]. More recently, a kernel was developed with an accuracy of 70% by using ADMET features, guided random walk generator and heterogeneous similarity based on SMILES and SMARTS strings [16].

Thus, from the literature alluded to above, it is understood that most of the existing research adopts different similarity measures, and their results show improvements in precision but not in recall. In this paper, the similarity-based ensemble prediction model is developed to identify the potential DDIs. It uses Jaccard's coefficient for similarity measures, and the neighbor recommender method and random walk method for the prediction of DDIs. The base prediction model is further improved by genetic algorithm techniques. The ensemble model developed identifies the drug-drug interactions for their different feature types. The experimental results shows that the random walk method along with the genetic algorithm improves precision, recall and thereby accuracy also. The approach is also tried with DDIs between approved and unapproved drugs, to highlight how this work can be used to understand DDIs arising in new drugs that may be clinically administered to patients who are already likely to be taking certain other particular drugs for treatment of related conditions. This in turn could save on the costs and efforts of clinical trials, and also help direct such trials to verify the occurrence—or lack thereof—of DDIs in specific drug-pairs of interest.

The rest of the paper is organized as follows: Section II describes the data sets used. Section III explains the basic approach, Section IV discusses the results and Section V concludes the results and describes possible applications of this model.

II. Dataset Description

Five different databases are used. They are:

- PubChem (https://pubchem.ncbi.nlm.nih.gov/): this is a public repository for information on chemical substances and their biological activities. It contains 93.9 million chemical compounds [17]
- DrugBank (https://www.drugbank.ca/): this combines detailed drug data with comprehensive drug target and drug action information. It contains 11,682 drug entries [18]
- SIDER (http://sideeffects.embl.de/): this combines data on drugs, targets and side effects into a more complete picture of the therapeutic mechanism of actions of drugs and the ways in which they cause adverse reactions. It contains 1430 drugs, 5880 adverse drug reactions (ADR) and 1,40,064 drug-ADR pairs [19]
- KEGG (https://www.genome.jp/kegg/drug/): this is a comprehensive drug information resource for approved drugs in Japan, USA, and Europe, unified based on the chemical structure and/or the chemical component, and associated with therapeutic target, metabolizing enzyme, and other molecular interaction network information.
- Offsides (http://tatonettilab.org/resources/tatonetti-stm.html): this contains side effects for 1332 drugs and 10,097 adverse events [20]

We created a combined dataset using features from the above five databases (using PHP scripts).[‡]

Drug data can be broadly classified into three types - chemical, biological and phenotypic.

- Chemical data consists of substructures data. It is obtained from the PubChem database. There are around 881 different substructures.
- Biological data consists of targets, transporters, enzymes, and pathways. Target, transporter and enzyme data are obtained from the DrugBank database, and pathways data are obtained from the KEGG database. There are 780 types of targets, 78 types of transporters, 129 enzyme types, and 233 pathway types.

[‡]See https://figshare.com/articles/dataset_ddiPrediction/7454312 for our data.

- Phenotypic data consists of indications, side effects, and off-side effects. Indications and side effects data are obtained from the SIDER database and off-side effects data are obtained from the Offsides database. There are 4897 types of indications and side effects, 9496 off-side effects.

Every drug in the database has a feature called *substructure* which is associated with an array of size 881, such that the value of the array at a given position is 1 if the corresponding substructure exists in the drug, and it is 0 otherwise. This representation is used for all the other features of every drug. The dataset is described in Table I.

TABLE I
FEATURE TYPES, FEATURES AND DATABASE

Feature Type	Feature	Database	Number of Types
Chemical	Substructure	PubChem	881
Biological	Targets	Drugbank	780
Biological	Transporters	Drugbank	78
Biological	Enzymes	Drugbank	129
Biological	Pathways	KEGG	253
Phenotypic	Indications	SIDER	4897
Phenotypic	SideEffect	SIDER	4897
Phenotypic	OffSideEffect	Offsides	9496

III. METHODOLOGY

This section describes the proposed ensemble model as shown in Figure 1. The dataset, drug substructure data, drug target data, drug enzyme data, drug transporter data, drug pathway data, drug indication data, drug side effect data, drug offside effect data and known drug-drug interactions, as shown in Table I, have been used. Multi-source data provide biological information, chemical information, phenotypic information, and known interactions to characterize drug-drug interactions.

The dataset has been prepared and their features are identified as previously indicated. The fact that similar drugs can potentially lead to drug-drug interactions is applied here. Because we rely on similarity of features of drugs (for potential DDIs), we use a similarity index to calculate the similarity between two drugs. Each drug has an array of size 881 corresponding to its substructure feature. There are different kinds of similarity measures [21], like Euclidean distance, Manhattan distance, and Jaccard's coefficient. We use Jaccard's coefficient for this model. Jaccard's coefficient best captures how many common substructures two drugs share and how many features the two of them do not have. This describes similarity more accurately than calculating the distance between the two feature values.

The Jaccard's coefficient similarity index is calculated for each of the eight features shown in Table I. These similarity values are used in two representative methods: neighbor recommender method and random walk method, to build a DDI prediction model. These models with ensemble rules (the weighted average ensemble rule and the classifier ensemble rule) are integrated and developed into an ensemble model to achieve better performance.

A genetic algorithm uses the objective function to maximize the AUPR (area under precision-recall curve) score, and logistic regression is used for the classification. These base prediction models and the role of the genetic algorithm in the ensemble model are further described in subsequent subsections.

A. Similarity Measure

Consider two drugs X and Y with vectors V_X, V_Y corresponding to a particular feature. The similarity is then calculated using Jaccard's formula, as shown in (1):

$$S(V_X, V_Y) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{1}$$

where M_{11} is the number of positions in V_X and V_Y, where both have a value of 1, M_{01} is the number of positions in which V_X has a value 0, and V_Y has a value 1, M_{10} is the number of positions in which V_X has a value 1 and V_Y has a value 0. Eight similarity values are obtained, one for each of the eight features.

B. Neighbor Recommender Method

The neighbor recommender method is one of the most popular methods in recommender systems [22], which recommends items (movies, music, books, etc.) to users, or predicts the rating or preference that users would give to items. This method is intuitive and relatively simple to implement. In its simplest form, only one parameter, i.e., the number of neighbors used in the prediction, requires tuning. This method provides a concise and intuitive justification for the computed predictions.

There are many advantages for this method when compared with other methods:

(i) One of the strong points of neighborhood-based system is its efficiency. Unlike most model-based systems, this method does not require costly training phases which need to be carried at frequent intervals in large commercial applications. It may require pre-computing to find the nearest neighbors in an offline step, but this is definitely much cheaper than model training and re-training.

(ii) This method provides near-instantaneous recommendations.

(iii) Moreover, storing these nearest neighbors requires very little memory, making such approaches scalable to applications with millions of users and items.

(iv) Another useful property of recommender systems based on this approach is that they are little affected by the constant addition and deletion of users, items and ratings, which are typically observed in large commercial applications. For instance, once item similarities have been computed, an item-based system can readily make recommendations to new users, without having to retrain the system. Moreover, once a few ratings have been

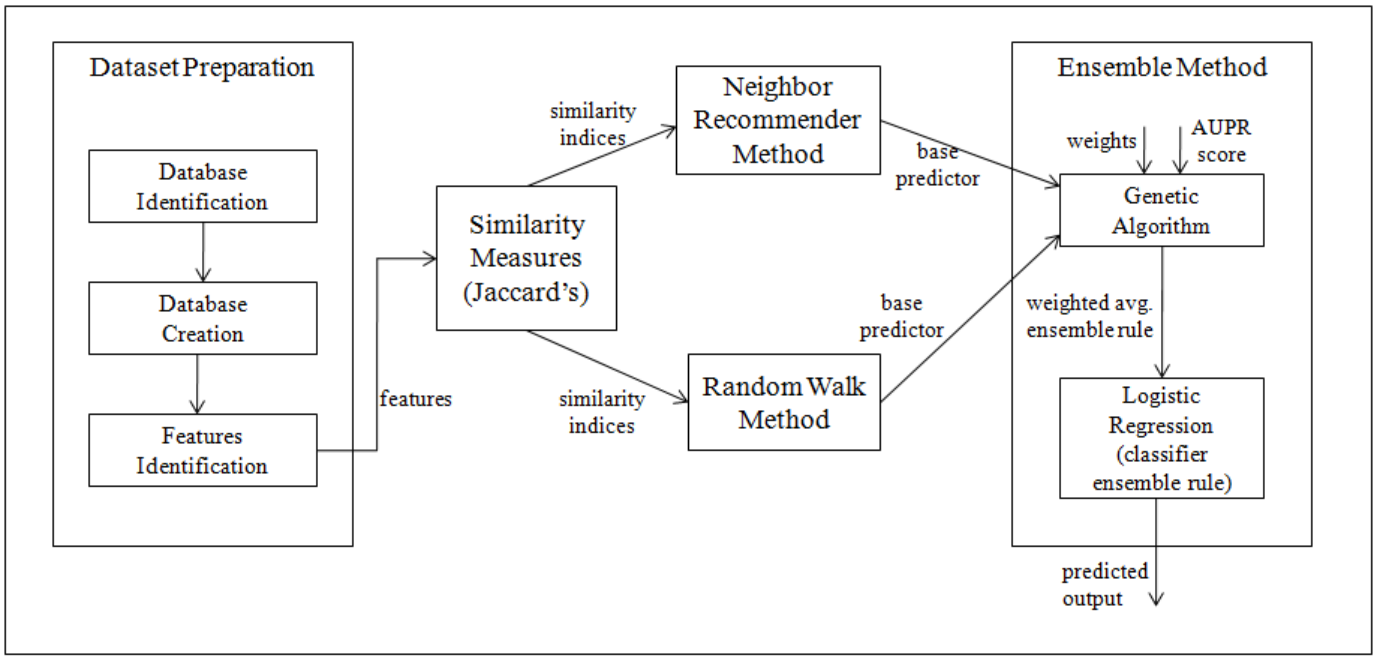

Fig. 1. Framework of Ensemble Model

entered for a new item, only the similarities between this item and the ones already in the system need to be computed. These reasons explain why neighborhood recommender method is chosen as one of the similarity-based methods for DDI prediction.

Given an $N \times N$ similarity matrix $S = (s_{ij})$ for N drugs, known pair-wise DDIs are denoted by an adjacency matrix $A = (a_{ij})$. The neighbor recommender method takes the weighted-average information of neighbors for prediction, as shown in (2). Y_{ij} is calculated for drug_i and drug_j, as mentioned below with $a_{kj} = 0$ when drug_k and drug_j do not cause any DDI and $a_{kj} = 1$ when drug_k and drug_j cause a DDI.

$$Y_{ij} = \frac{\sum_{k=1,k\neq j}^{N} s_{ik} a_{kj}}{\sum_{k=1,k\neq j}^{N} s_{ik}} \tag{2}$$

Similarly Y_{ji} is calculated using (2). The probability that drug_i interacts with drug_j is then given by $Y_{ij} + Y_{ji}$. This value is calculated for each of the eight features.

Algorithm 1: Neighbor Recommender Method

Data: $N \times N$ similarity matrix $S = (s_{ij})$ for N drugs
Result: Probability of drug_i interacting with drug_j
Adjacency matrix $A = (a_{ij})$

$$Y_{ij} = \frac{\sum_{k=1,k\neq j}^{N} s_{ik} a_{kj}}{\sum_{k=1,k\neq j}^{N} s_{ik}}$$

Probability(drug_i interacts with drug_j) = $Y_{ij} + Y_{ji}$

C. Random Walk Method

In a random walk for any graph G, the sequence of points are selected randomly. That is we first pick a starting point and then randomly choose a neighbor of this point (say point A). We now move to this point A and then we do the same again. A random walk is a finite Markov chain that is time-reversible. In fact, there is not much difference between the theory of random walks on graphs and the theory of finite Markov chains; every Markov chain can be viewed as random walk on a directed graph, if weighted edges are used. Similarly, time-reversible Markov chains can be viewed as random walks on undirected graphs, and symmetric Markov chains, as random walks on regular symmetric graphs. In this method, a random walker starts from an initial node, and moves to neighbors with the probability μ and moves back to the initial node with the probability $1 - \mu$.

Given a $N \times N$ similarity matrix $S = (s_{ij})$ for N drugs, known pair-wise DDIs are denoted by an adjacency matrix $A = (a_{ij})$. This matrix A describes known DDIs i.e; $a_{ij} = 1$ if $drug_i$ and $drug_j$ interact, otherwise $a_{ij} = 0$. The similarity matrix S is normalized as shown in (3).

$$W = D^{-1}S \tag{3}$$

where D is the degree matrix of S. The matrix form of the update is summarized as shown in (4).

$$Y = \mu WY + (1 - \mu)A \tag{4}$$

This converges to the solution, as shown in (5).

$$Y = (1 - \mu)(1 - \mu W)^{-1}A \tag{5}$$

The probability that drug_i interacts with drug_j is then $Y_{ij} + Y_{ji}$. This value is calculated for each of the eight features.

Algorithm 2: Random Walk Method

Data: $N \times N$ similarity matrix $S = (s_{ij})$ for N drugs
Result: Probability of drug_i interacting with drug_j
Adjacency matrix $A = (a_{ij})$ Normalized S =

$$W = D^{-1}S$$

$$Y = \mu WY + (1-\mu)A$$

The converged solution is,

$$Y = (1-\mu)(1-\mu W)^{-1}A$$

Probability(drug_i interacts with drug_j) = $Y_{ij} + Y_{ji}$

D. Ensemble Method

It is quite natural to combine different prediction models for better performance. Ensemble learning is a useful technique that aggregates multiple machine learning models to achieve overall high prediction accuracy as well as good generalization [23]. Ensemble learning has in particular been applied to a great number of applications in bioinformatics [24].

An ensemble learning system usually has two components: base predictors and ensemble rules. In our ensemble system, the heterogeneous models $\{f_i\}_{i=1}^n$ based on multi-source data is adopted and they act as base predictors. To integrate base predictors, two popular ensemble rules are considered: the weighted average ensemble rule and the classifier ensemble rule. The weighted average ensemble rule takes the weighted average of outputs from the base predictor. For a new input x_{new}, base predictors give the predictions $\{f_i x_{\text{new}}\}_{i=1}^n$ and their weighted average $\sum_{i=1}^n w_i f_i(x_{\text{new}})$ is adopted as the prediction of the ensemble model, where $\sum_{i=1}^n w_i = 1$ and $w_i \geq 0$.

The genetic algorithm (GA) is used to determine weights in the ensemble model. In the GA optimization, candidate weights are represented as chromosomes, and the fitness of a chromosome is the area under the precision-recall curve (AUPR) score of the ensemble model on the validation data. The objective function of GA optimization is to maximize the AUPR score.

The classifier ensemble rule is to seek a classification function

$$G : (f_1(x), \ldots, f_n(x)) \rightarrow \{0, 1\}$$

which maps outputs of n base predictors to a label. For a new input x_{new}, outputs of base predictors are $\{f_i x_{\text{new}}\}_{i=1}^n$ and the prediction of the classifier ensemble model is $G(f_1(x_{\text{new}}), \ldots, f_n(x_{\text{new}}))$. The logistic regression is used as the classification function.

IV. Results and Discussion

A. Model Evaluation using k-fold cross validation

k-fold cross validation [25], [26] has been used to evaluate the predicted models. The data are randomly split into k subsets of equal size. In each run, one of these subsets is used as the testing data and the rest are used as training data and the model has been built. In order to avoid bias of data-split, 20 runs have been done for k-fold cross validations, and the accuracy values are shown in Table II.

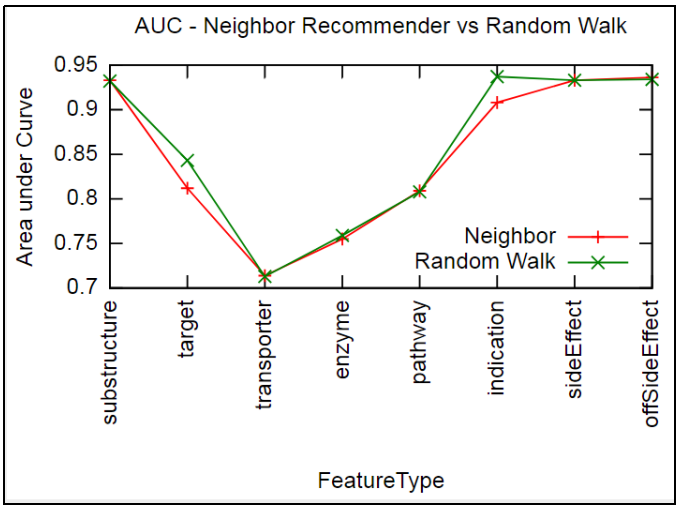

Fig. 2. Area under Curve for Prediction Models

TABLE II
Accuracy - Performance of Similarity Models

Feature	Neighbor Recommender	Random Walk
Substructure	0.947	0.947
Target	0.863	0.926
Transporter	0.874	0.867
Enzyme	0.908	0.928
Pathway	0.929	0.936
Indication	0.919	0.950
SideEffect	0.946	0.947
OffSsideEffect	0.946	0.946

It is seen that the DDI is well determined for the features *Target* and *Indication* by the random walk method, with accuracy values 0.926 and 0.950 respectively, as shown in Table II.

The ensemble model is evaluated with the measures: area under curve (AUC) and area under precision recall (AUPR) curve. It is evident from Figure 2 that the AUC values by the random walk method are closer to 1, when compared with the neighbor recommender method. (The closer the AUC value to 1, the better the model is.)

The performance of random walk method for the prediction of drug-drug interaction is further visualized in Figure 3. It shows that the AUPR value for the features *Target*, *Enzymes* and *Indicator* are higher when compared with the neighbor recommender method. (The higher the AUPR value, the better the model is.)

From the Figures 4, 5 and 6, it is evident that the random

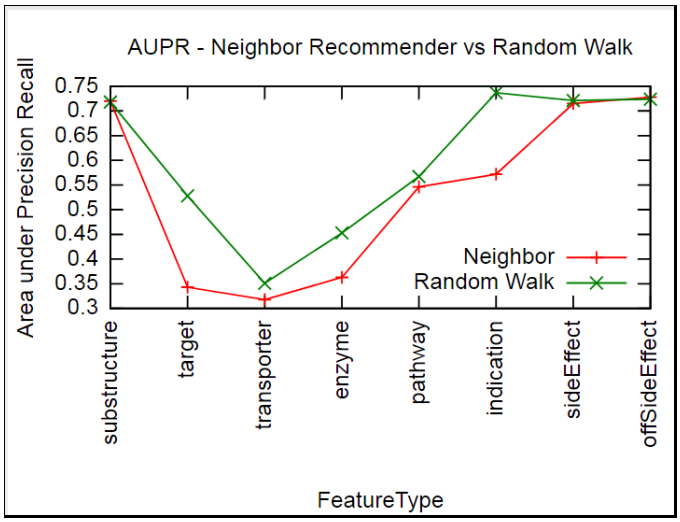

Fig. 3. Area under Precision Recall for Prediction Models

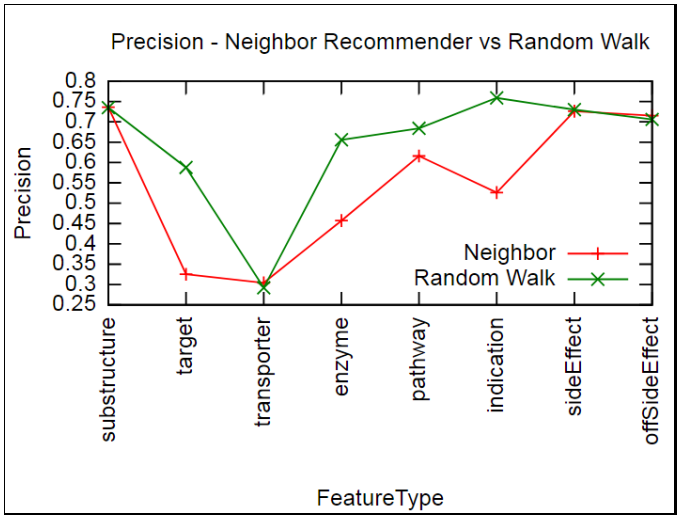

Fig. 4. Precision Values for Prediction Models

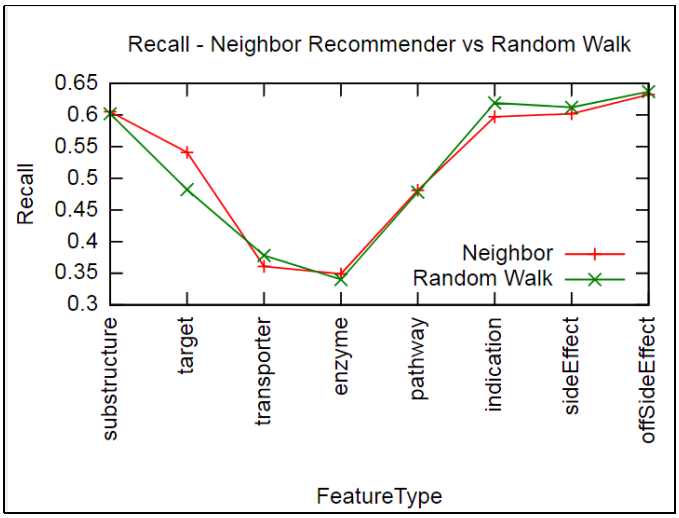

Fig. 5. Recall Values for Prediction Models

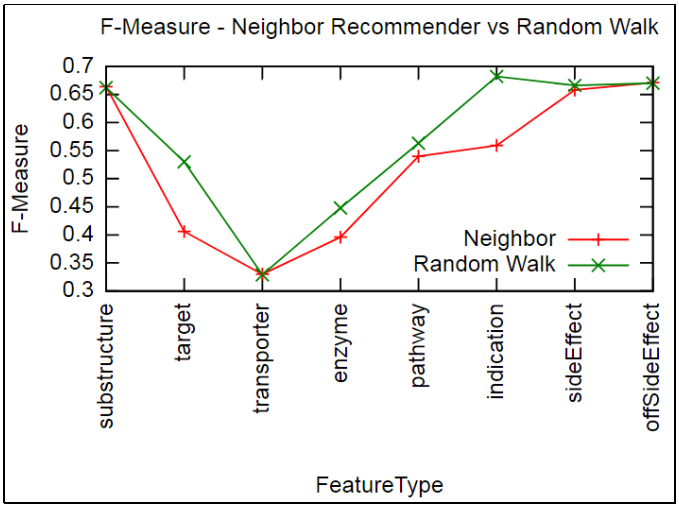

Fig. 6. F-Measure Values for Prediction Models

walk method as the base predictor model performs well when compared with the neighbor recommender method. The drug-drug interactions for most of the feature types are correctly predicted and it is known from the precision/recall/F-measure values.

From Figure 4, it is seen that the precision value of the feature *Target* is 0.6 by the random walk method, while it is 0.34 by the neighbor recommender method. There exist substantial differences in the precision values between these two methods for most of the features.

Similarly, the recall values by random walk method are greater for most of the features when compared with neighbor recommender method, as shown in Figure 5. This reflects in a better F-measure values by the random walk method as shown in Figure 6.

B. Results on approved drugs

Seven drug-drug pairs from the set of drug-drug interactions have been removed and the model is re-trained on the new subset of DDIs. Our model has been used to check DDIs on these seven drug-drug pairs. The results obtained are described in Table III. Here, 0 represents the case when two drugs do not interact and 1 represents the case when the two drugs do interact.

C. Results on unapproved drugs

Four unapproved drugs are identified in the dataset and the model is used to check whether they interact with some of the approved drugs. The interactions detected by the model are mentioned in Table IV. Here, 0 represents that two drugs do not interact and 1 represents that the two drugs interact.

TABLE III
RESULTS ON APPROVED DRUG-PAIRS

Drug 1	Name	Drug 2	Name	Expected Output	Observed Output
DB00001	Lepirudin	DB00002	Cetuximab	0	0
DB00001	Lepirudin	DB00248	Cabergoline	1	0
DB00005	Etanercept	DB00051	Adalimumab	1	1
DB00006	Bivalirudin	DB00029	Anistreplase	1	1
DB00009	Alteplase	DB00005	Etanercept	0	0
DB00001	Lepirudin	DB01323	St. John's Wort	1	1
DB00102	Becaplermin	DB00013	Urokinase	1	1

TABLE IV
RESULTS ON UNAPPROVED DRUGS

Unapproved Drug	Approved Drug	Name	Output
Denileukin diftitox	DB00072	Trastuzumab	1
Reteplase	DB06271	Sulodexide	1
Peginterferon alfa-2a	DB00811	Ribavirin	1
Leuprolide	DB00308	Ibutilide	0

V. CONCLUSION

Drug-drug interactions are of the greatest concern for any patients who need to take multiple drugs, and also for their physicians, care-givers, and society at large. Machine learning and related techniques have a significant role to play in helping obtain knowledge of possible DDIs, given the concerns with expense, effort, and ethics of relying on clinical testing alone.

The detection and the knowledge gained about such interactions using machine learning would enable the pharmaceutical industry to obviate the need for some testing, and would enable physicians to give provide better care and avoid adverse reactions.

In a clinical context, this model could be used by physicians to check if the drugs they want to prescribe interact with any of the drugs that the patient might already be using.

This model could also be used by FDA in their drug approval process by checking for potential DDIs.

Our results show that the proposed ensemble model for the prediction of potential drug-drug interactions gives a significant accuracy of greater than 90%. It is possible because of the use of Jaccard's coefficient similarity measure and the random walk method for the prediction model, which is further enhanced by GA techniques. The model also predicts the interactions with approved and unapproved drug pairs.

ACKNOWLEDGMENT

This work was supported in part by Amazon Web Services through an AWS Machine Learning Research Award.

REFERENCES

[1] N. C. for Health Statistics, "Health, United States 2014: With special feature on Adults Aged 55–64. Hyattsville , MD ," 2015. [Online]. Available: https://www.cdc.gov/nchs/data/hus/hus14.pdf

[2] "Research c for de and drug interactions and labeling — Preventable Adverse Drug Reactions: A Focus on Drug Interactions." [Online]. Available: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/\DevelopmentResources/Dru

[3] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC Bioinformatics*, vol. 18, no. 18, 2017.

[4] M. Takarabe, D. Shigemizu, M. Kotera, S. Goto, and M. Kanehisa, "Network-based analysis and characterization of adverse drug-drug interactions," *J Chem Inf Model*, vol. 51, no. 11, pp. 2977–85, 2011.

[5] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, and C. Friedman, "Drugdrug interaction through molecular structure similarity analysis," *J Am Med Inform Assoc*, vol. 19, no. 6, p. 1066 –74, 2012.

[6] S. Vilar, E. Uriarte, L. Santana, N. Tatonetti, and C. Friedman, "Detection of drug-drug interactions by modeling interaction profile fingerprints," *PLos One*, vol. 8, no. 3, p. e58321, 2013.

[7] A. Gottlieb, G. Stein, Y. Oron, E. Ruppin, and R. Sharan, "Indi: a computational framework for inferring drug interactions and their associated recommendations," *Mol Syst Biol*, vol. 8, p. 592, Jul. 2012.

[8] A. Cami, S. Manzi, A. Arnold, and B. R. BY, "Pharmacointeraction network models predict unknown drug-drug interactions," *PLos One*, vol. 8, no. 4, p. e61468, 2013.

[9] J. Huang, C. Niu, C. Green, L. Yang, H. Mei, and J. Han, "Systematic prediction of pharmacodynamic drug-drug interactions through protein-proteininteraction network," *PLoS Comput Biol*, vol. 9, no. 3, p. e1002998, 2013.

[10] F. Cheng and Z. Zhao, "Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties," *J Am Med Inform Assoc*, vol. 21, no. e2, pp. e278–86, 2014.

[11] K. Park, D. Kim, S. Ha, and D. Lee, "Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks," *PLos One*, vol. 10, no. 10, p. e0140816, 2015.

[12] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Label propagation prediction of drug-drug interactions based on clinical side effects," *Sci Rep*, vol. 5, p. 12339, 2015.

[13] P. Li, C. Huang, Y. Fu, J. Wang, Z. Wu, J. Ru, C. Zheng, Z. Guo, X. Chen, and W. Zhou, "Large-scale exploration and analysis of drug combinations," *Bioinformatics*, vol. 31, no. 12, pp. 2007–16, 2015.

[14] A. Fokoue, O. Hassanzadeh, M. Sadoghi, and P. Zhang, "Predicting drug-drug interactions through similarity-based link prediction over web data," in *WWW'16 Companion, Proceedings of the 25th International Conference Companion on World Wide Web*, Montréal, Québec, Canada, 2016, pp. 175–178.

[15] R. Ferdousi, R. Safdari, and Y. Omidi, "Computational prediction of drug-drug interactions based on drugs functional similarities," *Journal of Biomedical Informatics*, vol. 70, pp. 54–64, 2017.

[16] D. S. Dhami, G. Kunapuli, M. Das, D. Page, and S. Natarajan, "Drug-drug interaction discovery: Kernel learning from heterogeneous similarities," in *Third IEEE/ACM Conference on Connected Health: Applications, Systems, Engineering and Technologies (CHASE'18)*, sep 2018, pp. 88–100.

[17] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "Pubchem substance and compound databases," *Nucleic Acid Research*, vol. 44, no. Database issue, p. D1202–D1213, 2015.

[18] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acid Research*, vol. 36, no. Database issue, p. D901–D906, 2007.

[19] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The sider database of drugs and side effects," *Nucleic Acid Research*, vol. 44, no. Database issue, p. D1075–D1079, 2015.

[20] N. Tatonetti, G. F. GH, and R. Altman, "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports," *J Am Med Inform Assoc*, vol. 19, no. 1, pp. 79–85, 2012.

[21] F. G. Ashby and D. M. Ennis, "Similarity measures," *Scholarpedia*, vol. 2, no. 12, 2007, doi:10.4249/scholarpedia.4116.

[22] A. A. Deshmukh, P. Nair, and S. Rao, "A scalable clustering algorithm for serendipity in recommender systems," in *International Conference on Data Mining Workshops (ICDMW 2018)*, Nov. 2018.

[23] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, Third Quarter 2006.

[24] P. Yan, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, 2010.

[25] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, 1974.

[26] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, no. 2, Apr. 2011.