Machine learning drugs side effects prediction

Zoran Gavrilov and Ana Madevska Bogdanova
Faculty of Computer Science and Engineering,
"Ss. Cyril and Methodius" University Skopje, R.N. Macedonia

zorangavrilov@yahoo.com
ana.madevska.bogdanova@finki.ukim.mk

***Abstract*— Adverse drug reactions can be the cause of hospitalization, increased morbidity and mortality, withdrawal of drugs from the market and consequently increased costs of the healthcare system. Current methods for predicting and assessing potential side effects are challenging in terms of costs and efficiency. Machine learning could be implemented for predicting the side effects of drugs. Therefore, we present machine learning classifier for predicting drugs side effects using different supervised learning models on a dataset consisted of chemical, biological and phenotypic features. Compared to other machine learning models for prediction of side effects of drugs, our model has similar and comparable performance. Machine learning probably wouldn't be able to predict all side effects, but it could help scientists to notice potential problems early and develop safer drugs in the future.**

Keywords—machine learning, side effects, supervised learning

I. INTRODUCTION

An adverse drug reaction is an unwanted reaction caused by the administration of a drug. Adverse reactions can cause hospitalization, increased morbidity and mortality, and increased costs of the healthcare system [1]. Identification of the adverse effects is of exceptional interest and importance in the drug discovery today. Potential drugs are tested on a small subset of the population that does not always represent the genetic and physiological composition of the general population. This causes side effects to occur in some ethnic groups even after the medicinal product has been approved by regulatory authorities. A large number of approved drugs are withdrawn from the market due to side effects. Therefore, recognizing potential side effects would help to reduce the costs and to avoid risks in drug discovery. The main method for predicting or assessing potential side effects early in drug development is the application of preclinical in vitro safety profiling by testing compounds with biochemical and cellular assays. However, such experimental detection remains a challenge in terms of costs and efficiency [2].

Data science is a rapidly growing field that can be used for analytical purposes, i.e. risk recognition and assessment, and prediction, i.e. machine learning that can predict events based on historical data. Data science can be applied or is already applied in various aspects of healthcare such as drug discovery, diagnosis, personalized treatment, etc [3]. Traditional computational methods analyze the structure-activity relationship or the quantitative structure-property relationship, but they are not suitable for large-scale data [4]. Therefore, machine learning methodology, which connects several machine learning methods have been proposed to predict side effects, there is still place for improvements. In this paper we present machine learning classifier for predicting drugs side effects using different supervised learning models.

II. METHODOLOGY

The model is built in Google Colab, which is a product of Google Research and allows writing and executing Python code through the browser and using the scikit learn library. One of the main and most important aspects of machine learning is data. There are several public databases for drugs, drug side effects, and similar related information. SIDER database contains information about drugs approved on the market and their adverse reactions [5]. PubChem database contains chemical information, i.e. information about the chemical structure of drugs [6,7]. DrugBank database combines detailed information about drugs with comprehensive information about drug targets [8,9]. Chemical structure of the drug is usually considered the most important factor for the side effects of the drug. Drug targets are typically involved in a specific metabolic or signaling pathway and can provide an important clue to drug side effects. Liu's data set was used to build the classifier model for predicting side effects of drugs, considering that it contains different characteristics (chemical structure, enzymes, signaling pathways, targets, transporters, indications, etc.) [10]. This dataset was compiled using the databases PubChem (from which the chemical structures were collected), DrugBank and KEGG (from which the biological properties were collected) and SIDER (for the phenotypic data). Each drug is associated with 1385 side effects represented as a binary dimensional profile, y(target variable), whose elements correspond to the presence or absence of each of the side effects with 1 or 0, respectively. Each drug is associated with three types of characteristics: chemical, biological, and phenotypic properties.

Type of feature	Specific feature	Source	Dimension
Chemical	Substructure	PubChem	881
Biological	Targets	DrugBank	786
	Transporters	DrugBank	72
	Enzymes	DrugBank	111
	Pathways	KEGG	173
Phenotypic	Indications	SIDER	869

Table 1. Dataset features and their subsets

Similarly to the target variable, the features are presented as a binary dimensional profile as well, ie with 0 and 1 corresponding to absence or presence respectively. In other words, the prediction of side effects is a binary classification problem where each drug causes or does not cause a certain side effect.

The prediction of side effects is done using 4 different models: support vector machine, logistic regression, decision tree and K-nearest neighbor which are also compared to each other. Each of these models is built for the 5 most common side effects: dyspnea, insomnia, thrombocytopenia, anorexia and paresthesia.

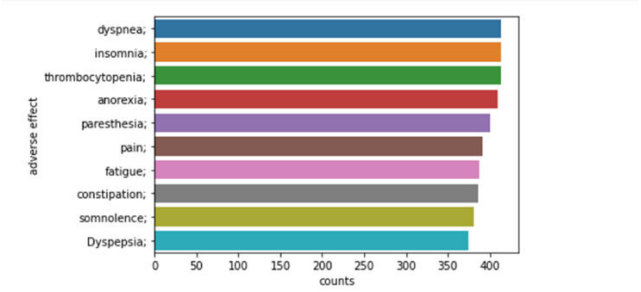

Figure 1. Most common side effects in the dataset

III. RESULTS

The essence of model evaluation is to determine the model's accuracy on potential unknown/out-of-sample data. One of the most important aspects in tackling the machine learning challenge is determining how effective the models are. Machine learning algorithms are evaluated using various performance metrics such as Log-Loss, Accuracy, AUC, Recall, Precision and F1 score for the classification models. The results from evaluation of the models are summarized in tables 2,3,4 and 5, while mean results of the models are summarized in table 6.

IV. DISCUSSION

Our model is built on Liu's dataset [10] and compared to this model, it has similar performance, i.e. similar values for precision and recall, especially for the support vector machine (0.675 and 0.722 respectively) and logistic regression model (0.5321 and 0.691 respectively).

Support Vector Machine							
Side effect	precision		recall		f1 score		accuracy
	0	1	0	1	0	1	
dyspnea	0.52	0.71	0.71	0.52	0.60	0.60	0.60
insomnia	0.64	0.62	0.59	0.67	0.62	0.65	0.63
thrombocytopenia	0.65	0.74	0.76	0.62	0.70	0.68	0.69
anorexia	0.57	0.74	0.72	0.60	0.64	0.66	0.65
paresthesia	0.67	0.77	0.82	0.59	0.74	0.67	0.70

Table 2. Results from evaluation of the Support Vector Machine model

KNN							
Side effect	precision		recall		f1 score		accuracy
	0	1	0	1	0	1	
dyspnea	0.53	0.53	0.53	0.53	0.53	0.53	0.67
insomnia	0.63	0.66	0.70	0.59	0.66	0.62	0.64
thrombocytopenia	0.66	0.68	0.74	0.59	0.70	0.63	0.67
anorexia	0.51	0.61	0.60	0.52	0.55	0.56	0.55
paresthesia	0.65	0.59	0.76	0.46	0.70	0.52	0.63

Table 3. Results from evaluation of the KNN model

Decision Tree							
Side effect	precision		recall		f1 score		accuracy
	0	1	0	1	0	1	
dyspnea	0.58	0.64	0.81	0.36	0.68	0.46	0.60
insomnia	0.62	0.53	0.36	0.76	0.46	0.62	0.55
thrombocytopenia	0.56	0.52	0.47	0.61	0.51	0.56	0.61
anorexia	0.61	0.69	0.81	0.45	0.70	0.55	0.64
paresthesia	0.61	0.57	0.57	0.61	0.59	0.59	0.63

Table 4. Results from evaluation of the Decision Tree m

Logistic regression										
Side effect	precision		recall		f1 score		ROC AUC score	Jaccard score	Log loss	accuracy
	0	1	0	1	0	1				
dyspnea	0.69	0.69	0.73	0.65	0.71	0.67	0.72	0.53	0.78	0.69
insomnia	0.61	0.64	0.67	0.57	0.64	0.60	0.69	0.45	0.71	0.62
thrombocytopenia	0.69	0.72	0.77	0.64	0.73	0.68	0.75	0.54	0.75	0.70
anorexia	0.59	0.72	0.72	0.59	0.65	0.65	0.72	0.48	0.77	0.65
paresthesia	0.76	0.64	0.71	0.70	0.73	0.67	0.74	0.54	0.66	0.70

Table 5. Results from evaluation of the Logistic Regression model

Zhang et al. developed a new method, Feature selection-based multi-label k-nearest neighbor method, which can simultaneously determine the critical dimensions of features and predict with multiple labels with high accuracy [4]. The average of the recall scores for 309 drugs is 0.463, and 0.609 for the top 200 drugs, which is also not very different from our model. MEDICASCY is a random forest machine learning model for predicting side effects of drugs based only on their chemical structure with an accuracy of about 78% as in [11]. Pauwels' model for predicting side effects of drugs is based only on their chemical structure [12] while Mizutani's model uses target proteins [13].

Based on the evaluation, the model proposed here has a performance that is similar or even better in certain aspects than these two models (accuracy).

V. CONCLUSION

From the four proposed models, the logistic regression model has the best results. But in general, all models need optimization and improvement. While machine learning probably wouldn't be able to predict all side effects, it could help scientists to notice potential problems early and develop safer drugs in the future.

More research is needed in machine learning for the prediction of side effects of drugs since there are not many studies on this issue.

Model	precision		recall		f1 score		accuracy
	0	1	0	1	0	1	
Logistic regression	0.67	0.68	0.72	0.63	0.69	0.65	0.67
KNN	0.60	0.61	0.67	0.54	0.63	0.57	0.63
Support vector machine	0.61	0.72	0.72	0.60	0.66	0.65	0.65
Decision tree	0.60	0.59	0.60	0.56	0.59	0.56	0.61

Table 6. Mean results

VI. REFERENCES

[1] C.Y. Lee, and Y.P.P Chen, ”Machine learning on adverse drug reactions for pharmacovigilance,” in Drug Discovery Today, vol. 24, issue 7, July 2019, pp. 1332-1343.

[2] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban, “Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development,” in Drug Discovery Today, vol. 10, Issue 21, November 2005, pp. 1421-1433.

[3] K. Szymanski, Prediction of Adverse Drug Reaction (ADR) Outcomes with use of Machine Learning. Feed-Forward Artificial Neural Network with Backpropagation (master thesis). Green, 2020.

[4] W. Zhang, F. Liu, L. Luo and J. Zhang, “ Predicting drug side effects by multi-label learning and ensemble learning,” in BMC Bioinformatics, vol. 16, November 2015, pp.365.

[5] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, and P. Bork, “A side effect resource to capture phenotypic effects of drugs,” in Mol Syst Biol., vol. 6, issue 1, 2010.

[6] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, and S.H. Bryant, “PubChem: a public information system for analyzing bioactivities of small molecules,” in Nucleic Acids Research, vol. 37, Jul 2009 , pp. W623-633.

[7] Q. Li, T. Cheng, Y. Wang, and S.H. Bryant, “PubChem as a public resource for drug discovery,” in Drug Discovery Today, vol. 15, issues 23-24, December 2010, pp.1052–1057.

[8] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, et al., “DrugBank: a comprehensive resource for in silico drug discovery and exploration,” in Nucleic Acids Res., 34(Database issue), January 2006, pp. D668-672.

[9] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, D. Cheng, S. Shrivastava, D. Tzur, et al., “DrugBank: a knowledgebase for drugs, drug actions and drug targets,” in Nucleic Acids Res., 36(Database issue), January 2008, pp. D901-906.

[10] M. Liu , Y. Wu, Y. Chen, J. Sun, Z. Zhao, X. Chen, et al., “Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs,” in Journal of the American Medical Informatics Association, vol.19, Jun 2012, pp.28–35.

[11] H. Zhou, H. Cao, L. Matyunina, M. Shelby, L. Cassels, J.F. McDonald, et al., “MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action,” in Mol Pharmaceutics, vol. 17(5), May 2020, pp. 1558–1574.

[12] E. Pauwels, V. Stoven, and Y. Yamanishi, “Predicting drug side-effect profiles: a chemical fragment-based approach,” in BMC Bioinformatics, vol. 12, May 2011, pp.169.

[13] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi, “Relating drug-protein interaction network with drug side effects,” in Bioinformatics, vol. 28, issue 18, September 2012pp. 522–528.