ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
March 2020, Vol 13(09), 1046 – 1056
DOI: 10.17485/ijst/2020/v13i09/148136,

Drugs Classifier System Based on Machine Learning Algorithms

Abdulaziz Shehab[1,2]*, Kamal Al dayah[2] and Ibrahim Elhenway[3]

[1]Department of Computer Science, College of Science and Arts, Jouf University, Kingdom of Saudi Arabia
[2]Department of Information Systems, Faculty of Computers and information, Mansoura University, Egypt
[3]Department of Computer Science, Faculty of Computers and information, Zagazig University, Egypt

Article Type: Article

Article Citation: Abdulaziz Shehab, Kamal Al dayah, Ibrahim Elhenway. A drugs classifier system based on machine learning algorithms. *Indian Journal of Science and Technology*. 2020; 13(09), 1046-1056. DOI: 10.17485/ijst/2020/v013i09/148136

Received date: October 14, 2019

Accepted date: February 10, 2020

***Author for correspondence:**
Abdulaziz Shehab ✉ aishehab@ju.edu.sa 📍 Department of Computer Science, College of Science and Arts, Jouf University, Kingdom of Saudi Arabia

Abstract

Background/objectives: Nowadays, there are thousands of approved drugs that can be used for treating people who have medical problems. Therefore, drug warnings and precautions are denoted to recognize a discrete set of adverse effects and other implied protection uncertainties that are useful for patient control. **Methods/analysis/findings:** In this study, the intended framework is divided into two principal stages: data retrieval and data processing. Firstly, in the data collection stage, drug reports, drug interactions, malfunctions, number of deaths, and other factors had been obtained from various references, including RxNorm and Drug Bank using web service. Secondly, in the data processing phase, different data mining algorithms used to classify drugs into suitable drugs and non-suitable drugs. **Application/improvements:** According to the experimental results, we found that the decision tree has more accuracy (97.9%) than other state-of-art methods.

Keywords: Drug Interactions, Drugs Classification, Naïve Bayes, Support Vector Machine, Decision Tree.

1. Introduction

Usually, several data mining methods have been utilized in healthcare, such as classification, association, analysis, clustering, and regression, as shown in Figure 1. A short explanation of each one of them is presented next.

1.1. Classification Techniques

Classification breaks data units into distinct groups. The categorization procedure foretells the aim class for several data points. For example, patients can be classified as "great danger" or "normal" patients with their illness model using data organization strategy. It is a supervised training procedure having identified class divisions. Binary and multilevel

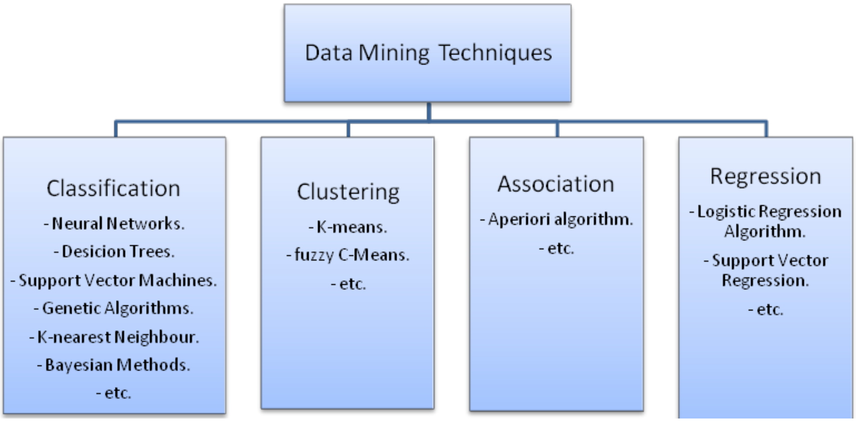

FIGURE 1. Different techniques used in the management of patient care.

are the two arrangements of classification. In a binary arrangement, only two available conditions such as, "true" or "false" danger inmate may be considered while the multiclass strategy has more than two purposes for example, "large," "moderate," and "fading" danger inmate [1–3].

Classification contains two footsteps. The initial step is designing structure, which is applied to explain the training records of a database. The additional step is designing method where the assembled model that used for classification. The efficiency of the categorization is assessed rendering to the rate of test units or test dataset that are properly classified [4, 5]. There is a comprehensive collection of different ways that have been used in healthcare supervision to complete the coordination process, which includes: J-48, SVM, K-nearest neighbor, neural networks, Bayesian methods, etc.

1.1.1. Decision Tree Algorithm

Decision Trees is a procedure very regularly used within data mining. The approach is to generate a collection of rules which can foretell a particular question variable based on a group of input data. A Decision Tree consists of vertices and edges. The edges express a way or a decision directing to the next vertices, maybe a pendant vertex (a pendant vertex is the vertices from which there are no additional edges to move), which could describe the subsequent question or statement. J48 is a public source Java implementation of the C4.5 method. The C4.5 method is an expansion of the ID3 algorithm and is utilized to initialize a Decision Trees that can be applied for classification.

1.1.2. Naive Bayesian Algorithm

Bayesian classification is utilized in data mining that can foretell the possibility of the class association. Bayesian classification is based on Bayes Theorem and is various regularly applied in machine learning. There are numerous distinct variants of Bayesian classification where Naive-Bayes is the most obvious.

1.1.3. Drug Interactions

Drug–drug interactions (DDIs) denote probable conflicting medication reactions having a significant influence on illness protection [6]. DDI is a situation when an individual drug affects another. The discovery of DDI is essential for both inmate protection and effective fitness administration [7]. DDIs are grouped into three principal types: no interaction, influence, and advice [8].

On the other hand, the adverse medication effect is a factor of the necessary forms of morbidity and destruction in the US, considering for above 700,000 crisis agency calls and 120,000 inmates annually. Adverse drug interactions (ADIs) have been referenced as dormant causes of illness morbidity as well as raised pharmaceutical costs and negligence cases [9]. Data Mining techniques in uncovering and inferring the hidden patterns from huge amounts of data in many fields, including the medical field encourage us to use it in inferring DDIs, the serious adverse reactions, and the clinically significant reactions associated with drugs.

Although there is a high estimate of medication datasets and semi-structured sources (e.g. Stockley [10–11]) with knowledge about DDIs, these datasets are inadequate, and the proportion of their information is restricted, so it is tough to select original clinical consequences to every interaction.

The main problem that discussed in this article is drugs recommendation and identify which efficient drugs than others. Therefore, medications should be recommended for the inmates correctly. However, physicians should able to classify drugs based on the knowing of drug details such as adverse effect, patient reports, drug alerts, and drug precautions which complicated task is due to a massive number of medications.

The primary contribution of this work can be summarized in two points. First, many drugs have serious adverse reactions, warnings, precautions, and other factors that can affect human lives or can cause severe medical problems. Therefore, it is very critical for physicians knowing the different warnings and precautions associated with each drug and can classify them to most suitable one during the drugs prescription process. Second, despite of the obvious importance of drug in prescribing decisions and patient management, there is currently no single complete source for drug warnings and precautions.

The remaining parts of this document are arranged as follows: Part 2 presents a summary of remarkable efforts that have been done for coordinating and interpreting the drug reports. Part 3 includes the recommended system with a comprehensive explanation of each step toward developing the recommended mechanism. Part 4 illustrates the implementation practice and evaluation that describe the suggested mechanism. Finally, conclusion will be introduced in part 5.

2. Previous Works

2.1. Drug to Drug Extraction and Classification Approaches

In this sub-section, we comprise some of the research works that have been performed in the field of DDIs extraction and classification. In Ref. [12], the aim to collect the scattering of drug information on the web among different databases that may entail incomplete drug guidance details.

Therefore, in this work, we aim to build drug interaction ontology containing information about adverse drug reactions and drug precautions, side effects and uses by integrating different drug resources.

In Ref. [13], the authors presented a new kernel-based features scheme to obtain and analyze drug interactions described in the biomedical literature. Like many previous works, their method consists of two steps. First, they detect interacting drug pairs, and then they classify each extracted pair into one of four interaction categories. Then, they used a binary classifier (LIBSVM classifier is used with RBF kernel) to detect interacting drug pairs. When evaluated on the DDIExtraction 2013 challenge corpus, their system achieved an F1-score of 71.14%.

In Ref. [14], the writers investigated the aggregate of pairing various machine-learning techniques to obtain DDI: (i) a feature-based approach adopting an SVM with a collection of attributes derived from texts, and (ii) a kernel-based approach mixing three different kernels. Investigations attended on the DDIExtraction2011 challenge corpus show that our method is helpful in selecting DDIs with 0.6398 F1 scores.

In Ref. [15], the writers included a scheme promoted to select DDI for drug specifying combinations observed in biomedical documents. This approach relies massively on deep syntactic parsing to represent the relationships among drug remarks.

In explaining the DDI extraction operation, they assessed the compatibility of both text-based and database obtained characteristics for DDI discovery. For machine learning, they examined both SVM and RLS approaches, with particular investigations for defining the optimal factors and training strategy. Their scheme has produced an achievement of 62.99% F-score on the DDI Extraction 2011 task.

In Ref. [16], the producers formed a corpus of Federal medicine Administration recommended medicine container supplement records that have been manually interpreted for pharmacokinetic DDIs by a pharmacologist and a medicine information specialist.

Then, they estimated three various machine learning algorithms (SVM, and J48) for their experience to 1) recognize pharmacokinetic DDIs in the package insert corpus and 2) analyze pharmacokinetic DDI records by their modality (i.e., whether they report a DDI or no interaction between medication pairs [17] (Table 1).

2.2. Web Services Concepts

In the proposed system, we heavily depended on web services to collect the domain knowledge. In this sub-section, a summary is provided on the basic concepts of web services. Web Services can be categorized into pair principal classes: SOAP API and REST API Web Services. The architectural style of this organization used in the implementation process.

SOAP is an OOP approach that determines a conventional rule applied for transferring XML-based information. It is illustrated as protocol designation for transferring structured data in the developing of Web Services in machine interfaces. The designation describes an XML-based case for moving information, and the protocol specifies a set of controls for transforming platform-specific data models into XML descriptions.

TABLE 1. Comparison between different related works

Reference	Year	Data	Method	Conclusion
In [19]	2012	1061 drugs, 172 context and 41 relations	Random Forest for DDI	Identification of DDI based on drug to drug relationship with accuracy 91%
In [20]	2011	FAERS of 37 drugs with adverse event profiles	Latent signal detection algorithm	Model for identify an adverse effect
In [21]	2013	Drug Bank DDI data	Logistic Regression Model and Apriori	Development of structured models and showed the best results in DDI with accuracy 95%
In [22]	2015	WHO Vigi Base of 2275 reported drugs	Naïve Bayes and Logistic Regression Model	Co-reported medications were associated with changes in liver event
In [23]	2015	FAERS and EMR data of 601 DDI with warfarin	Semantic web and ontology	Protentional DDIs with accuracy 92%
In [24]	2016	HER with 345 drugs and 10 adverse events	SVM and priorization of DDIs	Priorization of DDIs using four sources with accuracy 93%

Representational State Transfer (REST) means a source-oriented technique, and it signifies described by fielding in as a structural form that includes of a collection of scheme guidelines that determine the appropriate behavior for applying web patterns such as HTTP. Although REST is basically described in the circumstances of the web, it is becoming a popular implementation method for generating web services.

RESTful is developed with Web models (URI, HTTP, and XML) and REST sources. REST policies include connectivity, addressability, and stateless. RESTful applied to determine particular actions that operated on URL sources. Nevertheless, individually has its separate characteristics and weaknesses that make it more or less fitting for several kinds of application as given in Table 2.

TABLE 2. Comparison between two different web services techniques SOAP API and REST API

Factors	SOAP API	REST API
Type	Tightly	Loosely
Domain	single URL representing service	URL for every source
Protocols	ALL	HTTP
Caching	NO	YES
Interface	Non-uniform	Uniform
Data Types	Binary	ALL
Method	HTTP Request	HTTP Request
Expandability	NO	Yes
Standard	WSDL, UDDI	HTTP, XML and MIME

2.3. Drug Databases

Due to the critical impact of DDIs on inmate protection and health care cost, as a standard representation of conflicting results, substantial examination purposes have been achieved to realize DDI information. In this part, we satisfy the remarkable of the specific disciplines. Lately, there are many accessible datasets and semi-structured references that include drug reports, including DDIs knowledge, such as Drug-Bank, and RxNorm [18].

DrugBank is a wide-ranged online database which contains general pharmacological and biochemical data about medicines, which argues their tools of treatment and their objectives. It is produced, managed, and improved by comprehensive research studies expressed by field-special trained curators.

RxNorm offers a dictionary for normalized titles of clinical medications. It was initially developed to treat all designated medicines in the USA. It involves a specific drug's actual component, dosage, interactions, and strengths.

3. Proposed System

In this section, a block diagram of the proposed system based interactive tool is represented in Figure 2. We moved through two stages via the suggested framework called data retrieval and data processing. In the next sub-sections, two phases of the proposed system

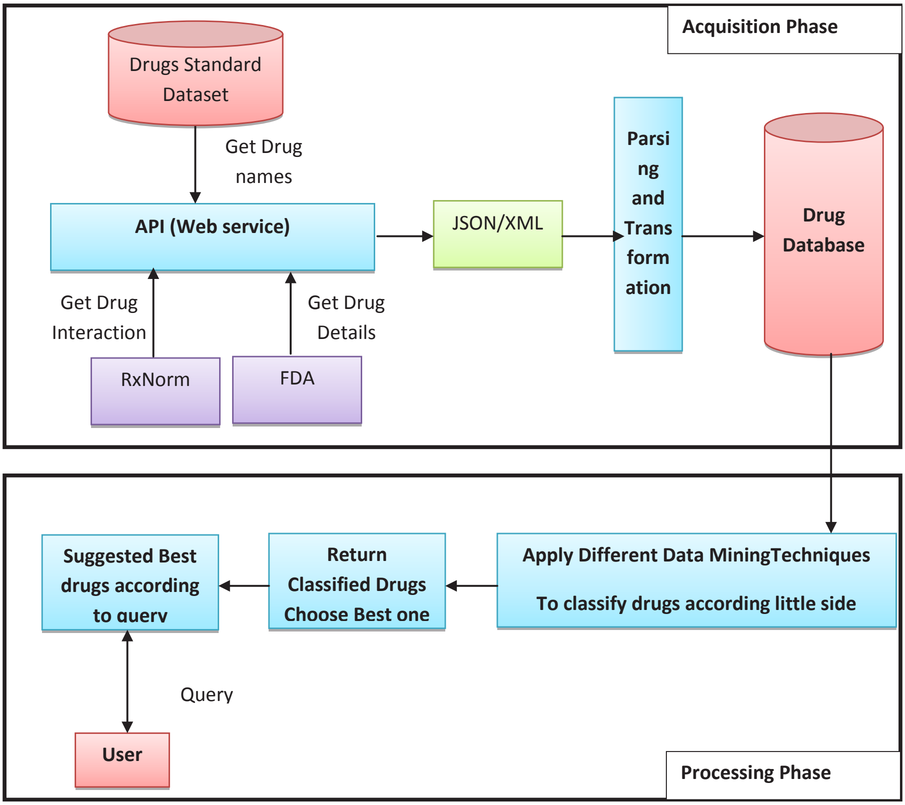
Acquisition Phase
Drugs Standard Dataset
Get Drug names
API (Web service)
JSON/XML
Parsi ng and Trans form ation
Drug Database
Get Drug Interaction
Get Drug Details
RxNorm
FDA
Suggested Best drugs according to query
Return Classified Drugs Choose Best one
Apply Different Data MiningTechniques To classify drugs according little side
Query
User
Processing Phase

FIGURE 2. The block structure of the proposed framework.

are explained in detail. For easier reading, the following statement is the definition of main two blocks in proposed system. Two phases of proposed system are drugs acquisition and data processing.

Acquisition Phase: in which the drug information is collected from different sources and stored in a relational database.

Processing Phase: in which we intend to build classification module using data mining techniques which classify drugs based on drugs information that collected from different resources.

3.1. Drug information Collection

Drug reports had been obtained from various origins, e.g., central drug dataset, RxNorm, and FDA. The retrieval manner of drug information begins with taking the drug titles from the drug primary standard dataset. After that, for several drug title, an HTTP request is sent to the RxNorm dataset is explored using web service to perceive the various drug titles and drug interactions.

Additionally, the drug anticipations, alerts, conflicting effects, evidence and, usage, are received from FDA, FDA includes documentation that suggested by Drug generators and suppliers about their stocks. It is essential for labeling comprises a review of the crucial scientific information necessary for the efficient and reliable use of the drug. The open FDA drugs stock labeling API presents data from this obedience for both delivered and over-the-counter drugs which are additionally broken down into segments, such as suggestions for use (prescription medications) or purpose, conflicting effects, and so on. HTTP Requests with URL using specifically to the drug labeling endpoint.

3.2. Information Processing

The next stage toward developing the recommended interactive tool, as shown in Figure 2, applies many methods for classification proper drugs such as Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and algorithms.

4. Datasets and Implementation Evolutions

The recommended interactive mechanism has been executed using some technologies, including WAMP server version 2.2, extensible markup language, PHP hypertext preprocessing scripting language, and MYSQL version 5.1. The dataset includes information about 468 drugs. The review of tools is represented in Table 3.

4.1. Result of Each Algorithmic Proposed System

4.1.1. Decision Tree

In this experiment, we conduct experiment using above tools and applied decision tree method on our dataset. We have 455 correctly classified patients out of 468 records with accuracy 97.9% in 0.28 second as shown in Table 4.

TABLE 3. Experiment tools

Internal server	Apache server
External server	FDA, RXNORM using HTTP and RESTful web service
Platform	Windows 7
Programming	PHP version 5.3
Data base engine	MYSQL version 5.5

TABLE 4. Decision Tree (J48) result

Relation name	Drugs_info
Number of attributes	10
Training time	0.28 seconds
Testing time	0.04
Tree size	25
Leave number	13
Correct classified data	455
In correct classified data	13
True positive	0.974
False positive	0.020
ROC	0.979

4.1.2. Random Forest

In this experiment, we conduct experiment using above tools and applied random forest method on our dataset. We have 453 correctly classified patients out of 468 records with accuracy 96.2% in 0.77 second as shown in Table 5.

TABLE 5. Random forest result

Relation name	Drugs_info
Number of attributes	10
Training time	0.77 seconds
Testing time	0.15 seconds
Iterations	100
Correct classified data	453
In correct classified data	15
True positive	0.972
False positive	0.024
ROC	0.962

4.1.3. Support Vector Machine

In this experiment, we conduct experiment using above tools and applied SVM method on our dataset. We have 288 correctly classified patients out of 468 records with accuracy 61.5% in 0.04 second as shown in Table 6.

4.1.4. Naïve Bayes

In this experiment, we conduct experiment using above tools and applied Naïve Bayes method on our dataset. We have 283 correctly classified patients out of 468 records with accuracy 60.2% in 0.12 second as shown in Table 6.

TABLE 6. SVM (support vector machine) result

Relation name	Drugs_info
Number of attributes	10
Training time	0.04 seconds
Testing time	0.03 seconds
Iterations	100
Correct classified data	288
In correct classified data	180
True positive	0.618
False positive	0.38
ROC	0.615

4.2. Performance Measurements

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \ \text{FN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. In our experiment, we implement different data mining algorithms such as DT, RF, SVM, and NB as shown in Tables 4–7, respectively. According to previous sub-section, we found that decision tree has more accuracy and reliable result (precision and recall) than other methods as shown in Table 8.

TABLE 7. Naive Bayes result

Relation name	Drugs_info
Number of attributes	10
Training time	0.12 seconds
Testing time	0.11 seconds
Correct classified data	283
In correct classified data	185
True positive	0.604
False positive	0.39
ROC	0.602

TABLE 8. Comparative study between different data mining methods

Methods	Correct classified instances	Precision	Recall	Accuracy	Time in second
Decision tree	455	98%	97%	97%	0.28
Random forest	453	96%	97%	96%	0.77
Support vector machine	288	40%	61%	61%	0.04
Naïve Bayes	283	71%	60%	60%	0.12

5. Summary and Conclusion

In this study, we present an interactive framework that promotes the fitting to find a suitable and safe medication to the inmate before entering the patient clinical information and his/her history medication according to some circumstances such as drug interaction, number of side effects, number of deaths, etc. Drugs datasets had been collected from Drug bank, FDA, and RxNorm using web service API. Moreover, we conducted experiments using different data mining methods. The decision tree achieves 98% in terms of precision and 97% in terms of both recall and accuracy. Therefore, according to our study, it outperforms random forest, SVM, and Naïve Bayes methods.

References

1. Pu L. e ToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacology and Toxicology*. 2019, 20(1), 2. https://bmcpharmacoltoxicol.biomedcentral.com/articles/10.1186/s40360-018-0282-6
2. Rezaee R. An evaluation of classification algorithms for prediction of drug interactions: Identification of the best algorithm. *International Journal of Pharmaceutical Investigation*. 2018, 8(2), 92–99. https://www.jpionline.org/index.php/ijpi/article/view/255
3. Vamathevan J. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. 2019, 1. https://www.nature.com/articles/s41573-019-0024-5
4. Tomar D, Agarwal S. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Biotechnology*. 2013, 5(5), 241–266. http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25
5. Patel S, Patel H. Survey of data mining techniques used in the healthcare domain. *International Journal of Information*. 2016, 6(1/2), 53–60. http://aircconline.com/ijist/V6N2/6216ijist06.pdf
6. Ontology for drug-drug interactions. https://www.researchgate.net/publication/286834043_An_ontology_for_drug-drug_interactions. Date accessed: 01/2013.
7. Chowdhury FM, Abacha AB, Lavelli A, Zweigenbaum P. Two different machine learning techniques for drug-drug interaction extraction. *Challenge Task on Drug-drug Interaction Extraction*. 2011, 19–26. https://www.semanticscholar.org/paper/Two-Different-Machine-Learning-Techniques-for-Chowdhury-Abacha/9998ab164023c2400c725d68d5971579bbb19008
8. Automated extraction and classification of drug-drug interactions from text. https://www.researchgate.net/publication/278030155_Automated_Extraction_and_Classification_of_Drug-Drug_Interactions_from_Text. Date accessed: 01/2013.

9. Goldberg RM, Mabee J, Chan L, Wong S. Drug-drug and drug-disease interactions in the ED: analysis of a high-risk population. *The American Journal of Emergency Medicine*. 1996, 14(5), 447–450. DOI: 10.1016/S0735-6757(96)90147-3.
10. Stockley's drug interactions. https://about.medicinescomplete.com/publication/stockleys-drug-interactions/. Date accessed: 2010.
11. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A. Drug Bank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*. 2013, 42, 1091–1097. DOI: 10.1093/nar/gkt1068
12. Amer S, Mahmoud H, El-Shishtawy T. A methodology for building integrated drug interaction ontology. *International Journal of Advancements in Computing Technology*. 2018, 10(2), 1–9. http://www.globalcis.org/ijact/ppl/IJACT3620PPL.pdf
13. Raihani A, Laachfoubi N. Extracting drug-drug interaction from biomedical text using a feature-based kernel. *Journal of Theoretical & Applied Information Technology*. 2016, 109–120. http://www.jatit.org/volumes/Vol92No1/14Vol92No1.pdf
14. Chowdhury FM, Abacha AB, Lavelli A, Zweigenbaum P. Two different machine learning techniques for drug-drug interaction extraction. Conference: proceedings of DDI extraction. 2011.
15. Björne J, Airola A, Pahikkala T, Salakoski T. Drug-drug interaction extraction from biomedical texts with SVM and RLS classifiers. *DDI Extraction*. 2016, 35–42. https://pdfs.semanticscholar.org/d641/472d03c05dd1ae3b98acfc86e8e768711622.pdf
16. Boyce RD, Gardner G, Harkema H. Using SVM and decision tree to identify pharmacokinetic drug-drug interactions. Conference: proceedings of the workshop on biomedical natural language processing. 2018.
17. Mumbaikar S, Padiya P. Web services based on SOAP and REST principles. *International Journal of Scientific and Research Publications*. 2013, 3(5), 1–4. http://www.ijsrp.org/research-paper-0513/ijsrp-p17115.pdf
18. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: Rx Norm at 6 years. *Journal of the American Medical Informatics Association*. 2011, 18(4), 441–448. DOI: 10.1136/amiajnl-2011-000116.
19. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. In: Pacific symposium on biocomputing. 2014, 410–421. https://www.ncbi.nlm.nih.gov/pubmed/22174296
20. Tatonetti NP, Denny JC, Murphy SN. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology & Therapeutics*. 2011, 90(1), 133–142. DOI: 10.1038/clpt.2011.83.
21. Yan S, Jiang X, Chen Y. Text mining driven drug-drug interaction detection. Proceedings IEEE international conference on bioinformatics and biomedical technology. 2013, 349–355. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4133978/
22. Suzuki A, Yuen NA, Ilic K. Comedications alter drug liver injury reporting frequently: data mining in the WHO VigiBase. *Regulatory Toxicology and Pharmacology*. 2015, 72(3), 481–490. DOI: 10.1186/s13040-015-0044-6.
23. Jinag G, Liu H, Solbrig HR. Mining severe drug-drug interaction adverse events using Semantic Web Technologies: case study. *Bio Data Mining*. 2015, 8, 12. DOI: 10.1186/s13040-015-0044-6.
24. Banda JM, Callahan A, Winnenbrug R. Feasibility of prioritizing drug-drug event associations found in electronic associations found in electronic health records. *Drug Safety*. 2016, 39(1), 45–57. DOI: 10.1007/s40264-015-0352-2.