ISSN No:-2456-2165

Multi-Label Long Short-Term Memory-Based Framework to Analyze Drug Functions from Biological Properties

Pranab Das
Assam, India

Abstract:- Drug function identification from the drug properties is important in drug discovery. Each year billions of dollars are spent on empirical testing of the drugs, which is costly, chemical wastage, and time-consuming. The computational experiments would help reduce drug discovery time and cost significantly. Most of the existing works have focused on single-label drug function identification. However, the capability of the drug's biological properties (transporter, target, carrier, and enzyme) has not yet been explored for multiple drug function identification. Identifying drug function is a multi-label classification problem. So, in the present work, a multi-label long short-term memory-based framework has been proposed for identifying drug function. The data related to biological properties has been extracted from DrugBank, and drug functions are collected from PubChem. The proposed framework performance has been found promising in terms of accuracy, precision, recall, F1, ROC-AUC score, and hamming-loss, and it achieved the highest accuracy of 95.80%.

Keywords:- *Multi-Label, LSTM, Biological Properties, Drug Function, Machine Learning.*

I. INTRODUCTION

Drug development is one of the essential procedures in pharmaceutical manufacturing production. Analyzing drug function is a vital part of the drug discovery, development, and design. The process of the drug development pipeline is a complicated, expansive, resource-consuming, chemical wastage, and time-needed process [1, 2]. There is a need to analyze drug function efficiently to avoid the maximum cost and time; hence, different computational methods are constantly being developed for analysing drug function. Computational methods are essential to minimize the time and cost during drug design and discovery [3, 4]. Long Short-Term Memory (LSTM) is a promising computational drug development approach for a new drug [5, 6, 7]. Several methods are applied in drug development experiments to get early information about the drug. LSTM techniques provide the various benefits that help in drug discovery and decision-making on high-quality data for well-specified questions. Recently, LSTM has demonstrated its usefulness in the drug discovery process.

The use of biological properties of the drugs is increasing to discover new drug-drug interaction [8, 9, 10] and side effects identification [11, 12, 13, 14]. However, biological properties are not yet utilized to analyze drug function. Such a study efficiently analyzes drug function from the biological properties of the drug. In the field of pharmacology, biotechnology, drug discovery, development, and design, analyzing the drug functions are essential in discovering new drugs efficiently. A drug can have multiple drug functions. Therefore classifying a drug into different drug functions is a multi-label task [15, 16]. The analysis of drug function can be carried out using a Multi-Label Long Short-Term Memory (MLLSTM). Unlike single label identification classification, the multi-label identification approach identify one or more drug functions at the same time. This work demonstrates how the multi-label long short-term memory approach is used on biological drug properties to analyze various drug functions derived from the medical subject heading (MeSH) [17]. The common problem in multi-label classification tasks is that it faces a class imbalance problem. A multi-label dataset with class imbalance is a complex problem, and the result may be affected. So, Multi-Label Synthetic Minority Over-Sampling Techniques (MLSMOTE) have been used to address the class imbalance issue [18].

This paper employs a multi-label long short-term memory framework on biological properties to analyze drug function. The literature survey shows that the multi-label analysis of drug function is addressed before using only a 2D chemical structure. However, analyzing multiple drug functions for specific drug-using biological properties has not been explored yet. This type of drug properties may use in drug development to analyze drug functions.

The main motivation in this work is to check whether drug functions are trained with the biological properties of the drug with a multi-label long short-term memory approach to analyze drug function efficiently.

The organization of the work is as follows: related work is summarized in section II. The architecture for the proposed framework and drug properties has been described in section III. In section IV, parameter values for classification models and experimental results have been presented. Finally, in section V, the outcome of the experimental analysis of drug function has been concluded.

ISSN No:-2456-2165

II. RELATED WORK

Meyer et al. [19] identify drug function by employing convolution neural networks on the 2D structure and a random forest classifier on the 1D structure. The authors collected functions of drugs, chemical 2D structure, and 1D structure from the PubChem website. Further, they identify single-label drug function for a particular drug from the chemical 1D structure. They also showed how multi-label classification performs to identify multiple drug functions for a particular drug from the 2D chemical structure of drugs. In [20], the authors employed a semi-supervised method named as Multicontrastive based on the 2D structure to identify the function of a drug. This approach achieves better class identification accuracy than the different existing semi-supervised methods. The authors collected the drug 2D structure from PubChem and DrugBank, and 12 drug functions from PubChem. For implementing their experiments, they used the ResNext model. In conclusion, the authors find their approach shows significantly better results than the other existing approach, such as Pi-model, VAT, MixMatch, and Pseudo-labeling. Aliper et al. [21] showed how deep neural networks and support vector machine classifiers were applied on large transcriptional response datasets (gene expression data) to analyze the drugs' pharmacological characteristics (drug functions). The authors use 12 drug functions and consider only those drugs that belong to only one drug function class. Further, they collected gene information for three cell lines for 6p78 drugs over PC-3, A549, and MCF-7 cell lines from the LINCS L1000 website to analyze drug functions. In their experiment, the deep neural network model performs better than the support vector machine.

In drug development, biological properties may use to identify function of a drug. These properties (transporter, target, carrier, and enzyme) are widely used in drug discovery. Most of the researchers used biological properties as input features to identify drug targets, drug-drug interactions, and adverse drug reactions. However, biological properties are not yet utilized as an input feature to a computational model to analyze drug function.

The literature survey shows that the multiple drug function identification for a specific drug has been addressed through the only 2D chemical structure. However, identifying more than one drug function for a particular drug at the same time using biological properties has not been utilized, which has established the principle of the work in this paper.

III. ARCHITECTURE OF THE PROPOSED FRAMEWORK

This section describes the problem statement, biological drug properties utilized to identify drug function, and the framework to solve the stated problem. Let Drug = {$Drug_1$, $Drug_2$, $Drug_3$, ..., $Drug_k$, ..., $Drug_m$}be the set of drugs, X = {transporter, target, carrier, and enzyme} be the set of features of drug properties, and Drug_Function = {$Function_1$, $Function_2$, $Function_3$, ..., $Function_l$, ..., $Function_n$} be the set of drug function where each $Function_l$ represent the drug function for a $Drug_k$ with drug features X. A drug $Drug_k$ can have multiple drug function at the same time. Therefore, classifying a drug into various drug function can be viewed as a multi-label drug function identification problem, Fig. 1 presents the representation of multi-label drug function for drug using their corresponding drug properties. Table- I shows the multi-drug function for a specific drug, whose PubChem CID is 134688985 (drug name: Hyoscyamine sulfate), which has three drug function; Cardiovascular (C), Central Nervous System (CNS), and Respiratory (R).Abbreviations and Acronyms

Table I: Example of multiple drug function for a drug.

PubChem CID	C	CNS	Dermatological	Urological	……	R
134688985	1	1	0	0	0	1

Hence, the multi-label identification task is essential to identify multiple drug function based on the biological properties of drug. The aim of the multi-label identification of drug function is to assign multiple labels (drug function) for a drug $Drug_k$, which input is related to a collection of drug features (X), and output is a set of possible Drug_Function.

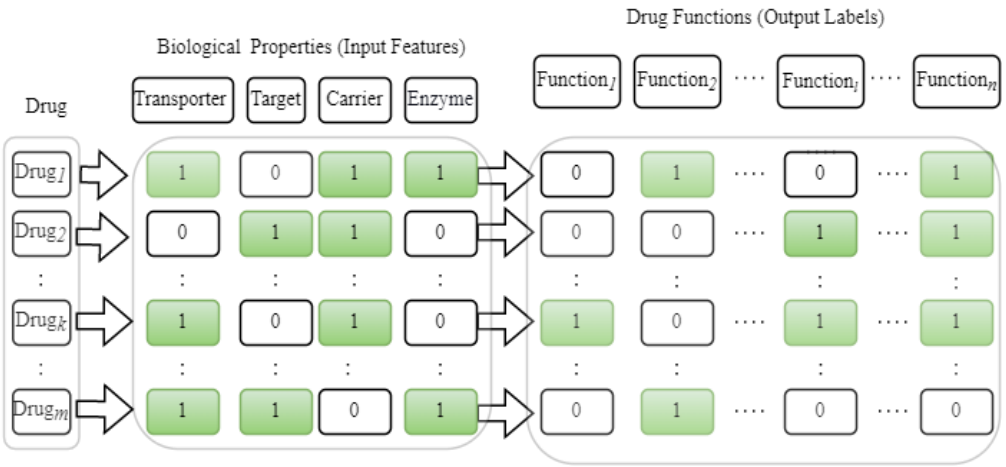

Fig. 1. Multi-label drug functions representation.

A. Dataset Description

In the proposed work, the drugs' biological information are used to identify drug function. The detailed description of the biological information and drug function are given below-

- *Drug Function*

Drug function is the adeptness of a specific drug (bioinformatics substance) to treat the targeted bodily part. These biochemical substances have been utilized to diagnose, cure, prevent or treat an ailment of any living tissue, which are the essential matters of the drug function. The drug function dataset contains a drug function with its corresponding PubChem CID extracted from PubChem [22]. Although PubChem consists of 20 high-level drug function, 12 drug function have been taken in this paper, described previously in Meyer et al. Drug function are represented with a well-ordered list of binary numbers 1 and 0 to indicate the presence and absence of drug function.

- *Biological Properties*

Biological properties are also crucial in silico experiments to drug discovery and development. In this work, transporter, target, carrier, and enzyme are used to classify drug function. The popular drug information database DrugBank [23] is used to retrieve the biological information. After mapping the drug biological properties with the drug function, it contains 1108 drugs corresponding to 12 drug functions.

The dataset with biological properties and drug function has been illustrated in Fig. 2, Where $Function_n$ is the total amount of drug function (n=12) and transporter, target, carrier, and enzyme are the properties of drugs.

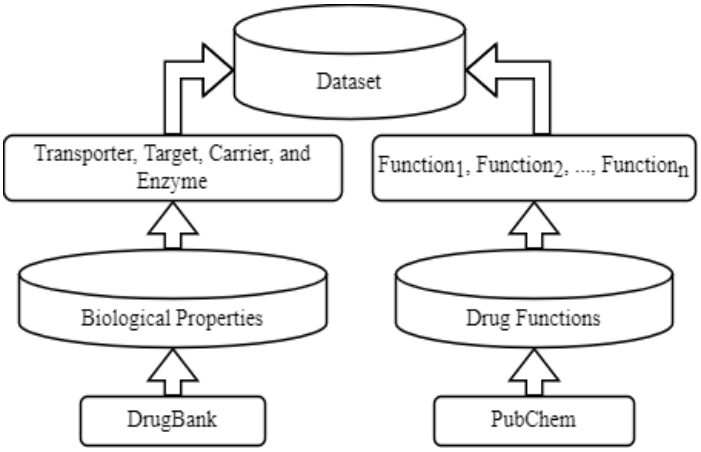

Fig. 2. Dataset preparation for the drug function identification.

B. MLSMOTE for Handling Class Imbalance

The popular and frequent problem in the Multi-label classification approach is unequal class distribution. When the number of class are not equal, class imbalance occurs in the dataset. The dataset with the frequency of class distribution has been illustrated in Fig. 3. In Multi-label classification learning, a dataset that has a class imbalance problem is a real-world obstacle complex problem that can cause result degradation. Dealing with this type of data is very important to get optimal results. So, a method named MLSMOTE has been taken. MLSMOTE algorithm assumes that a multi-label dataset may have one or more minority labels. In MLSMOTE, first, select the minority labels. Once a sample is selected which is belongs to minority labels, the MLSMOTE finds its nearest neighbor. After that, a set of synthetic sample features is generated by interpolation method.

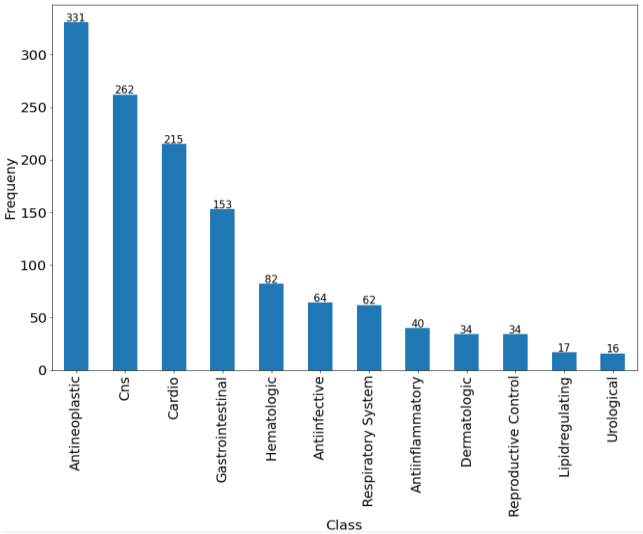

Fig. 3. Frequency of class distribution on protein dataset.

C. Proposed Methodology

For identifying drug functions, the input is related to biological features, and the output is the drug functions of a particular drug. One drug may have more than one drug function, so it belongs to the Multi-label task. A framework for the proposed methodology has been represented diagrammatically in Fig. 4. In the proposed methodology. For solving multi-label drug function identification task, a multi-label supported LSTM approach is proposed. Finally, MLLSTM classification algorithm performance is evaluated using different performance measures such as ROC-AUC, precision, hamming-loss, accuracy, recall, and f1 score.

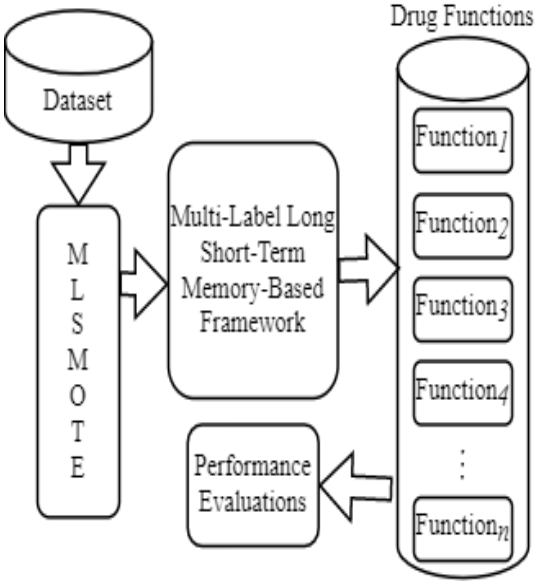

Fig 4: Work flow of identifying drug functions from biological properties.

ISSN No:-2456-2165

IV. PRAMETER VALUES FOR CLASSIFICATION MODELS AND EXPERIMENTAL RESULTS

A. Parameter Values for MLLSTM Model

A Multi-Label LSTM framework have been proposed to solve the multi-label drug function identification task. The proposed MLLSTM framework has been implemented in google colab using Python language (3.7.13 version). Keras and TensorFlow, with the help of sequential API is used to build the proposed model. The outcome of the MLLSTM framework is varied by the different number of hidden layer and units in each layer. The MLLSTM obtained better accuracy when the unit of the input layer was set to 16 and dropout 0.2 after that input and hidden layer. Further, it is observed that the performance of the MLLSTM does not improve as the number of hidden layer is increases. The MLLSTM model performs well on one hidden layer with 64 unit, and it achieved the highest accuracy. The performance of distinct units on each layer is shown in Table II. In the input layer return_sequence set to True, Adam is used as optimizer with binary_crossentropy as a loss function. The epoch is set to 10 and threshold is set to 0.5. If the probability of output is greater than 0.5, then the class label is assigned for that test sample; otherwise not. Learning rate 0.001 and tanh activation function is used for hidden layer, and sigmoid recurrent activation is set for output layer. The output layer neuron is set an equal number of labels (12 drug functions) and other parameters are set as default.

Input, Hidden Layer Sizes	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Hamming-Loss
16, 16	94.40 %	90.71%	87.34%	89%	97.59%	5.59%
16, 32	94.30%	91.88%	85.52%	85.55%	97.75%	5.66%
16, 64	**95.80%**	**92.05%**	**91.11%**	**91.60%**	**98.15%**	**4.23%**
32, 32	92.20%	91.70%	76.14%	83.20%	97.70%	7.82%
32, 64	90.40%	91.04%	69.28%	78.69%	97.46%	9.64%

Table II: Results of the MLLSTM approach on biological properties to identify drug functions.

B. Results

The outcomes of the experiment to identify drug function have been discussed in this section. The biological properties (transporter, target, carrier, and enzyme) were utilized to determine the drug function by employing a multi-label LSTM framework. The performance of the MLLSTM framework on biological properties has been presented in Table II.

It can be observed from Table II that the performance of the MLLSTM with input layer unit 16 and hidden layer unit 64 is comparatively better than the other hidden input layer unit. The MLLSTM model achieved the highest accuracy of 95.80%, precision score of 92.05%, recall value of 91.11%, F1 score of 91.60%, ROC-AUC score of 98.15%, and hamming-loss of 4.23%. The ROC-AUC score of the different hidden units of MLLSTM is of the proposed framework shown in Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9.

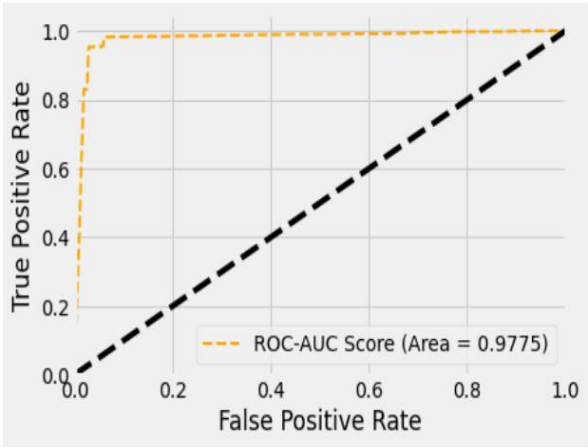

Fig. 6. ROC curve of MLLSTM classifier on biological properties for input and hidden layer unit 16 and 32 respectively.

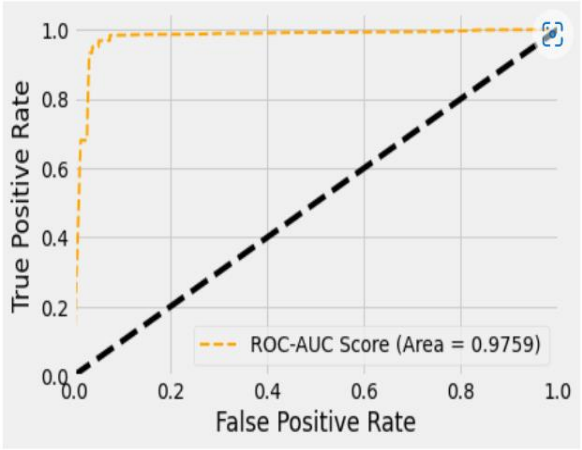

FIG. 5. ROC CURVE OF MLLSTM CLASSIFIER ON BIOLOGICAL PROPERTIES FOR INPUT AND HIDDEN LAYER UNIT 16 AND 16 RESPECTIVELY.

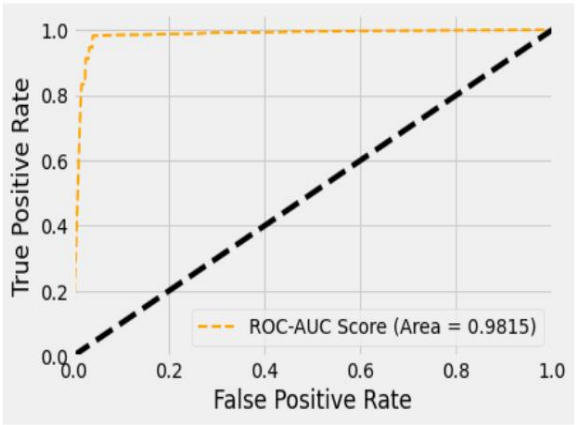

Fig. 7. ROC curve of MLLSTM classifier on biological properties for input and hidden layer unit 16 and 64 respectively.

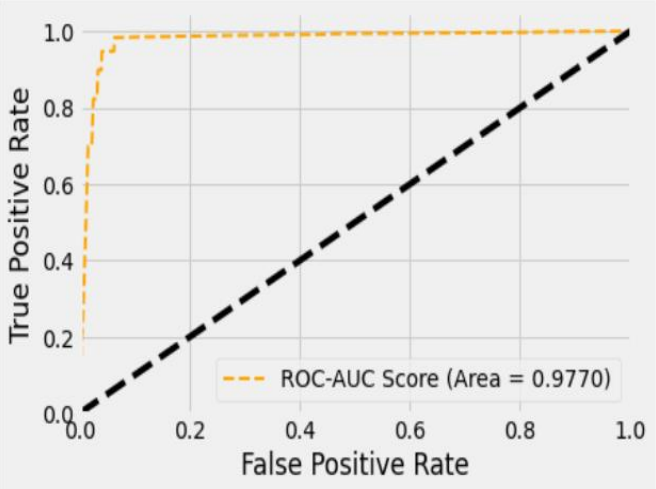

Fig. 8. ROC curve of MLLSTM classifier on biological properties for input and hidden layer unit 32 and 32 respectively.

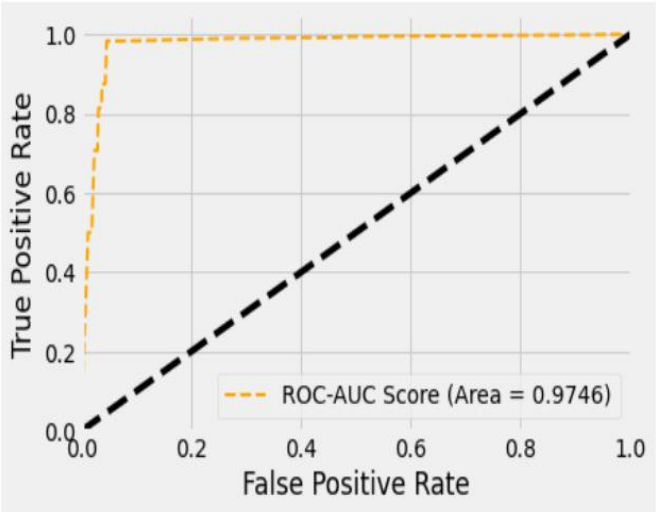

Fig. 9. ROC curve of MLLSTM classifier on biological properties for input and hidden layer unit 32 and 64 respectively.

V. CONCLUSION

The proposed methodology identifies drug functions by analyzing the biological properties of drugs by employing a multi-label long short-term memory-based framework. The drug function identification power of biological properties is sufficient. The proposed multi-label long short-term memory-based framework achieved the highest accuracy of 95.80% on the biological properties. Based on the achieved result, it can be said that the biological properties of the drug are essential for identifying the drug's function. Finally, this paper explores a multi-label long short-term memory-based approach to identifying multiple drug functions.

REFERENCES

[1]. Mohs, Richard C., and Nigel H. Greig. "Drug discovery and development: Role of basic biological research." *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 3.4 (2017): 651-657.
[2]. Taylor, David. "The pharmaceutical industry and the future of drug development." (2015): 1-33.
[3]. Hochreiter, Sepp, Guenter Klambauer, and Matthias Rarey. "Machine learning in drug discovery." *Journal of Chemical Information and Modeling* 58.9 (2018): 1723-1724.
[4]. Vamathevan, Jessica, et al. "Applications of machine learning in drug discovery and development." *Nature reviews Drug discovery* 18.6 (2019): 463-477.
[5]. Chen, Hongming, et al. "The rise of deep learning in drug discovery." *Drug discovery today* 23.6 (2018): 1241-1250.
[6]. Liu, Xiangyu, et al. "Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling." *International journal of clinical pharmacology and therapeutics* 59.2 (2021): 138.
[7]. Mouchlis, Varnavas D., et al. "Advances in de novo drug design: From conventional to machine learning methods." *International journal of molecular sciences* 22.4 (2021): 1676.
[8]. Ferdousi, Reza, Reza Safdari, and Yadollah Omidi. "Computational prediction of drug-drug interactions based on drugs functional similarities." *Journal of biomedical informatics* 70 (2017): 54-64.
[9]. Ibrahim, Heba, et al. "Similarity-based machine learning framework for predicting safety signals of adverse drug–drug interactions." *Informatics in Medicine Unlocked* 26 (2021): 100699.
[10]. Dere, Selma, and Serkan Ayvaz. "Prediction of drug–drug interactions by using profile fingerprint vectors and protein similarities." *Healthcare informatics research* 26.1 (2020): 42-49.
[11]. Liu, Mei, et al. "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs." *Journal of the American Medical Informatics Association* 19.e1 (2012): e28-e35.
[12]. Wang, Chi-Shiang, et al. "Detecting potential adverse drug reactions using a deep neural network model." *Journal of medical Internet research* 21.2 (2019): e11016.
[13]. Jamal, Salma, et al. "Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models." *Scientific reports* 7.1 (2017): 1-12.
[14]. Jamal, Salma, et al. "Computational models for the prediction of adverse cardiovascular drug reactions." *Journal of translational medicine* 17.1 (2019): 1-13.

[15]. Read, Jesse, et al. "Classifier chains for multi-label classification." *Machine learning* 85.3 (2011): 333-359.
[16]. Zhang, Min-Ling, et al. "Binary relevance for multi-label learning: an overview." *Frontiers of Computer Science* 12.2 (2018): 191-202.
[17]. Lowe, Henry J., and G. Octo Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches." *Jama* 271.14 (1994): 1103-1108.
[18]. Charte, Francisco, et al. "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation." *Knowledge-Based Systems* 89 (2015): 385-397.
[19]. Meyer, Jesse G., et al. "Learning drug functions from chemical structures with convolutional neural networks and random forests." *Journal of chemical information and modeling* 59.10 (2019): 4438-4449.
[20]. Sahoo, Pracheta, et al. "MultiCon: a semi-supervised approach for predicting drug function from chemical structure analysis." *Journal of Chemical Information and Modeling* 60.12 (2020): 5995-6006.
[21]. Aliper, Alexander, et al. "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data." *Molecular pharmaceutics* 13.7 (2016): 2524-2530.
[22]. Kim, Sunghwan, et al. "PubChem 2019 update: improved access to chemical data." *Nucleic acids research* 47.D1 (2019): D1102-D1109.
[23]. Wishart, David S., et al. "DrugBank 5.0: a major update to the DrugBank database for 2018." *Nucleic acids research* 46.D1 (2018): D1074-D1082.