DOI: 10.15849/IJASCA.220328.03

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 1, March 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Predicting Gene-Drug-Disease Interactions by integrating Heterogeneous Biological Data Through a Network Model

Hamza Hanafi, Badr Dine Rossi Hassani, and M'hamed Aït Kbir

LIST Laboratory UAE Tangier, Morocco
e-mail: hamzahanafi1@gmail.com
LABIPHABE Laboratory UAE Tangier, Morocco
e-mail: bd.rossi@fstt.ac.ma
LIST Laboratory UAE Tangier, Morocco
e-mail: maitkbir@uae.ac.ma

Abstract

Prediction of gene-drug-disease interactions have talented new insights in biology. Discovering unknown interactions will provide new therapeutic approaches to explore gene expressions. Recent improvements in machine learning techniques have gotten considerable interest due to higher efficiency, accurate results, and their lower cost. However, most of the studies were ignoring relevant associations, by representing only drug-disease interactions on a network while public available data offers a large variety of interactions. Additionally, some computational techniques used in this domain are faced with new challenges, related to the organization of heterogeneous data which suffer from a high imbalance rate since there are extensively more non-interacting gene-drug-disease triplets than interacting ones. In this paper we present integration of heterogeneous biological data about genes, drugs, and diseases to build a model, and building a new graph representation relating gene-drug-disease interactions. Using extreme gradient boosting (XGBoost) algorithm, we have been able to extract a list of valid interactions about gene-drug-disease triplets, and a list of gene-drug pairs related to lung cancer.

Keywords: *Biological heterogeneous data, Data integration, Gene-Drug-Disease interactions, Machine learning.*

1 Introduction

Traditional methods to develop new drugs are costly and time-consuming [1-4]. Computational techniques have gained increased interest to improve drug discovery. Nowadays, network analysis has revealed promising results in manipulating biological heterogeneous data. In addition, advanced new

Received 30 August 2021; Accepted 16 November 2021

technologies have generated a large amount of disparate data describing specific aspect of cells named Omics layers [5]. Using data integration methods along with network analysis have shown effective results to extract new interactions between biological data [6].

Data integration interests the creation of a model that combines data coming from different sources in order to explore new interactions more effectively. It has gained a lot of attention due to the large and different Omics datasets available. Several public databases provide the research community with a large amount of biological heterogeneous data which enabled to study biological processes and to support new findings in biology.

Network analysis are widely used in biology and extremely improved the exploration of relations among heterogeneous data. Biological systems are often represented as networks. They provide a mathematical representation of connections found in the literature. Consequently, they have become essential to understand biological mechanisms. Predictions are one of the most applied applications of network analysis, mainly to propose novel interactions. In this regard, machine learning (ML) algorithms have been widely used to build prediction models. Although, the efficiency of these techniques depends mostly on the training data and the preprocessing effort carried out over it.

Nowadays, numerous genes and drugs heterogeneous data is generated. This encourages the use of ML methods to learn from this data. One of the main difficulties with predictions of gene-drug interactions and gene-disease interactions is the volume of the data [7]. The use of an unbalanced dataset will result in an overfitting model. The number of single nucleotide polymorphisms (SNPs) present in the dataset highly affects the number of positive interactions, and genetic heterogeneity which may be common in complex diseases. Moreover, most of the studies do not present the preprocessing phase to handle unbalanced datasets such as feature selection [8].

Our contribution is to integrate heterogeneous data into two layers and to create a model capable of predicting gene-drug-disease interactions. We take full advantage of the disparate biological data present in DisGeNET and DGIdb databases to infer gene-disease interactions and gene-drug interactions. We constructed our dataset based on the inferred interactions and build a novel gene-drug-disease network. We learned from our data to predict valid interactions and evaluated three classifiers: XGBoost, ID3, and C4.5. We used the confusion matrix and individual feature contribution to evaluate the performance of the model. Our technique revealed effective and provided highly accurate prediction results.

The paper is organized as follows: In section 2 we present some related works. Our methodology is explained in section 3. The obtained results are discussed in section 4, followed by the conclusion.

2 Related Work

Statistical techniques conducted in predictions of biological interactions are founded on structure-based approaches or text mining methods. In structure-based approaches, the focus was put on the physical and chemical characteristics of the studied molecules. For example, prediction of gene-disease associations is based on genome-wide association study (GAWS); which selects a chromosome interval with candidate genes [9]. Similarly, prediction of drug-target interactions (DTIs) focuses on the binding drug sites. They require prior information on the binding sites. Consequently, they eliminate genes with unknown sequences. In text mining methods, the interactions between biological entities are inferred from the medical literature. Typically, this technique completely ignores unidentified interactions. Therefore, statistical methods have shown their limitation, besides they are time-consuming.

Computational methods have gained a lot of interest to improve prediction of biological interactions. Many methods have been proposed, similarity-based methods calculate similarity score between drug profiles to collect drug-drug interactions (DDIs). Vilar et. al [10] described numerous biological profiles that are used to compare the similarity. Drug structural profile is based on the fact that structurally similar drugs tend to target related genes [11]. Furthermore, similarity metrics have also been a subject of interest, Ferdousi et. al [12] compared several metrics and used the most optimal one to predict DDIs. The major disadvantage of these methods is to find a suitable threshold of similarity that is highly affected by false DDIs.

Networks-based methods have also been used to predict biological interactions, by constructing a network of interactions and then predict novel associations based on network analysis. In [13], the authors built a drug-drug similarities network based on several drug features and then used matrix factorization techniques to predict potential DDIs. Comparatively, in [14] the authors considered the interactions of DrugBank database to create an ensemble-based classifier using two techniques of matrix factorization: adjacency matrix factorization (AMF) and adjacency matrix factorization with propagation (AMFP). Other methods based on protein-protein interaction (PPI) networks covered a large number of DDIs detection. In [15] the authors used random walk algorithm to capture distant interference on PPI network. They assumed that DDIs are affected by close interference of triggered gene pathways. The PPI networks based-methods take full advantage of drug actions, but they suffer from incompleteness.

In the last few years, several ML methods have given more importance to data integration to improve the accuracy of predictions. Consequently, a substantial number of studies using different supervised learning approaches have been conducted. For example, in [16] the authors make use of a supervised manner to learn a kernel from heterogeneous similarities and different interaction types to predict DDIs. To evaluate their approach, they constructed a dataset from DrugBank database. Tong et al. [17] proposed a method to predict drug-target interactions using gradient boosting machines called SimBoost. The authors used features about drug-target (DT) pairs extracted from their network along with

similarity-based information. In a recent work on drug-gene interaction predictions, Zhu et al. [18] extended the metapath2vec and metapath2vec++ models into the gene-drug field and used both models on a biological heterogeneous network which involves three types of nodes: drugs, genes, and adverse drug reactions (ADRs). Actually, the two models can effectively represent the semantic of the heterogeneous information network. Even so, the precedent methodology was evaluated on a dataset comprising fewer ADR interactions than gene-drug pairs which may highly impact the accuracy of predictions. Most of the studies using ML methods require a feature engineering step. In fact, features used in the learning process highly affect the performance of the model. Nevertheless, current experiences do not present a comparative study of the used properties to reveal which characteristics are behind the prediction and which are less informative.

In this paper, we present a network-based approach to integrate heterogeneous biological data about genes, drugs, and diseases. We used an ensemble learning classifier, that has proven his superiority against ones usually used in the literature, to predict gene-drug-disease interactions. Furthermore, we present the feature engineering phase conducted in this study.

3 Methodology

Firstly, we describe the followed process to construct our dataset and to succeed the training phase. Gene-disease interactions are extracted from DisGeNET database. It is a discovery platform which stores human genes and variant-disease interactions. It also includes mendelian, rare, complex, and environmental diseases, as well as abnormal phenotypes and traits [19-21]. We selected 84038 curated gene-disease associations to constitute the gene-disease layer. Gene-drug interactions are collected from DGIdb database, a collection of various sources of gene-drug interactions as well as the druggable gene categories, 32107 gene-drug interactions are selected to constitute the gene-drug layer.

3.1 Graph Construction

A commonly used method to build an integrated graph is to project the edges of different graphs to the same set of nodes [6]. Hence, to construct our graph, we merged gene-disease interactions and gene-drug interactions into the same set of nodes. The created graph is named GDD (Gene-Drug-Disease). The overall framework is shown in Fig 1.

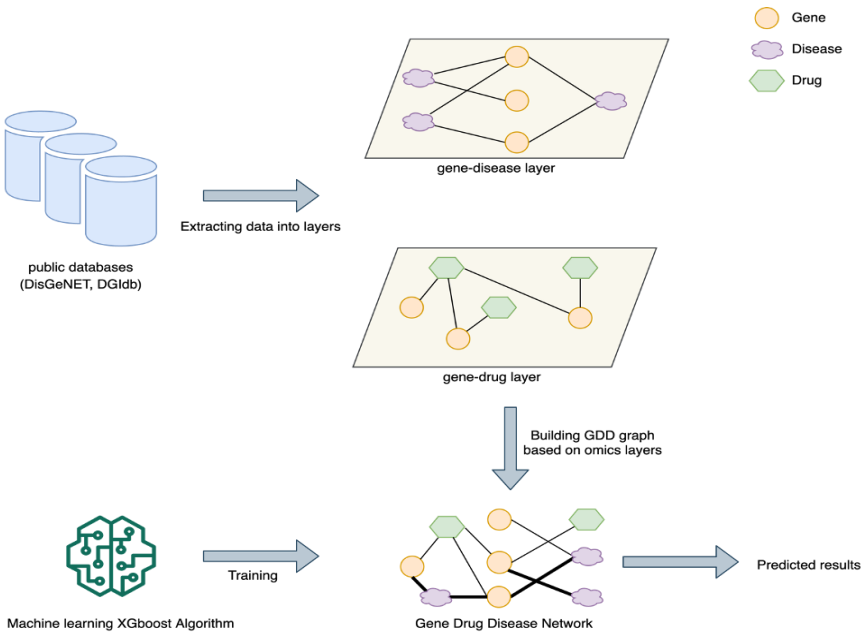

Fig 1. Overall framework we followed to build Gene-Drug-Disease interactions network and to learn from it.

We investigate several features to perform the training phase which is a supervised classification, where the target feature that we attempt to predict is the score value. The categorical features involved:

- Gene name;
- Disease name: name of the disease or the abnormal phenotype;
- Drug name;
- Evidence level (EL): a measure that denotes the strength of evidence of the interaction. It takes 6 values: strong, definitive, moderate, limited, disputed, and no reported evidence. Its value is computed by ClinGen.

The numerical features used:

- Evidence index (EI): a measure that shows the existence of paradoxical statements in articles talking about the same interaction;
- Disease specificity index (DSI): a measure that reveals if a gene is highly coupled with multiple diseases or only a limited number of diseases;
- Disease pleiotropy index (DPI): a measure that indicates if a set of diseases connected to a gene are similar among them;
- Probability of being loss of function intolerant (pLI): a gene metric that defines how much is a gene intolerant to loss-of-function variation (LoF variation). Its value is provided by the GNOMAD consortium;
- Score: defines the strength of the interactions in our GDD graph. It is a continuous value that ranges between 0 and 1, weak and strong associations respectively. its value is computed by DisGeNET database.

The values of the features are obtained from the DisGeNET database which computes them and stores them along with the interactions. The values for EI, DSI, and DPI are given by the following equations:

$$EI = \frac{\mathrm{N}_{pubs_{positive}}}{\mathrm{N}_{pubs_{total}}} \quad (1)$$

Here, $N_{pubs_{positive}}$ is the number of publications supporting the interaction and $N_{pubs_{total}}$ is the total number of publications.

$$DSI = \frac{\log_2(\frac{N_d}{N_T})}{\log_2(\frac{1}{N_T})} \quad (2)$$

Here N_d is the count of diseases linked to a gene and N_T is the total number of diseases.

$$DPI = \left(\frac{N_{dc}}{N_{TC}}\right) * 100 \quad (3)$$

Here, N_{dc} is the number of the various disease classes of the diseases related to the gene and N_{TC} is total number of disease classes in DisGeNET.

Our training set is constituted of 90% of the total samples and the 10% remaining samples is considered for the test set. The score feature is a continuous value that ranges between 0 and 1, this feature is mapped into 6 discrete values, after a resampling operation. the result is 6 target classes, denoted as follows: class 0, class 1, class 2, class 3, class 4, and class 5. Table 1 shows some data from the training set.

Table 1: Some data of the training set.

Gene Name	Disease Name	Drug Name	EI	EL	DPI	DSI	pLI	Score
BRAF	Neurofibromatosis 1	AEW-541	1	strong	0.79	0.35	0.9	0.38
CASR	Pancreatitis	ASP 7991	0.6	limited	0.65	0.47	0.060	0.46
NRAS	Noonan Syndrome	E-6201	1	definitive	0.69	0.42	0.52	0.77
ALMS1	Alstrom Syndrome	ZINC ION	0.96	definitive	0.62	0.5	0	1
ADA	melanoma	EHNA	1	no reported evidence	0.79	0.4	0	0.02

Decision tree-based algorithms are commonly used for classification problems. The goal is to predict a discrete value based on rules learned from the data features [22,23]. In our work, we compare several algorithms using trees to select the best association rule discovery method. We have compared the accuracy score of 3 algorithms: Iterative Dichotomiser 3 (ID3), C4.5, and XGBoost. Table 2 shows the results obtained for each algorithm.

Table 2: Accuracy score obtained for compared algorithms.

Algorithm	ID3	C4.5	XGBoost
Accuracy score	0.678	0. 706	0.898

XGBoost [24] clearly outperforms ID3 and C4.5 in term of accuracy score. It is built on top of universal gradient boosting methods with boosting capabilities to generate an ensemble of decision trees. Consequently, it turns out to be the best classifier among the compared algorithms.

4 Results, Analysis and Discussion

Our graph data is quite large. It represents more than 3 million gene-drug-disease interactions. To evaluate the performance of our classifier we created a sub dataset with samples randomly selected. The newly created sub dataset contains 10000 gene-drug-disease interactions. We used the Scikit-Learn implementation to train and evaluate our model. We used the grid-search function to obtain the optimal values for the hyperparameters of the model. The tuning we have performed achieved highest performance with 1000 boosted trees and a max depth of 7 level. To perform the k-folds cross validation technique our training data is split to k=10 folds [25]. This value was fixed experimentally, after using different values of k and comparing results.

Fig 2 shows the first 200 samples of the prediction results compared with their original values. Among them 7 were misclassified.

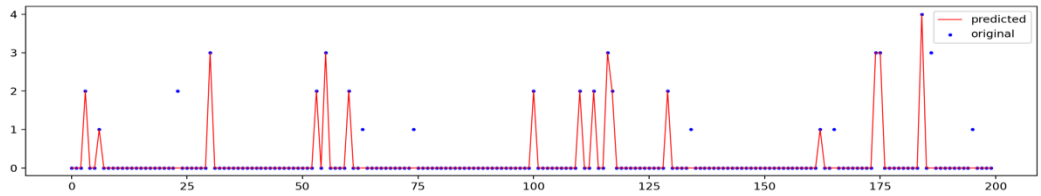

Fig 2. Predicted classes of the first 200 samples of our test set

Table 3 summarizes results of the confusion matrix. It is a way to visualize the capabilities of our ML model. It designates in its entries the count of predictions where the model correctly or incorrectly classified observations.

Table 3: Results of the 10-fold cross-validation of our sub dataset.

class	precision	recall	f1-score	count of samples
0	0.98	1.00	0.99	880
1	0.92	0.77	0.84	30
2	0.97	0.95	0.96	37
3	1.00	0.88	0.94	40
4	1.00	1.00	1.00	11
weighted avg	0.98	0.98	0.98	

Most samples present in the sub dataset are weak interactions and belong to class 0, strong interactions that belong to class 5 are not present because this class is minority. The recall value for class 1 is quite low; 77% compared to other classes, because our model is generating many false positive samples for class 1. This large unbalance in our data resulted to an overfitting (for each class). To overcame this drawback, we performed classical information gain to understand how the used features impact predicted results. There are three standard important measures to

explore in a model using trees. The first measure is based on the weight, it shows how many times a feature was selected to divide the data. The second measure is based on the cover, it's based on weighting the selected feature by the count of training samples going through those splits. The third measure is based on the gain, it presents the median training loss reduction gotten by selecting a feature to divide the samples. These measures remain global feature attribution and we need individualized explanation for each feature to assign feature importance.

The SHAP method [26] is a technique to find the mean variation in predictions taking into consideration selection of all possible features. Fig 3 shows individual feature contributions using SHAP method. It illustrates participation of each feature of the model, the rows show the impact of each feature on the predicted class.

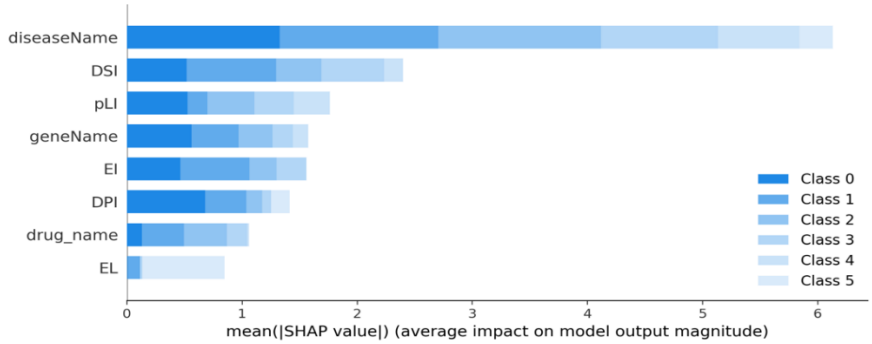

Fig 3. Individual feature contribution based on used sub-dataset, using six classes

Our experiments show that some features have weak impact on some classes or no impact at all. For example, the EI feature takes values that belong to only four classes. Unlike the disease name feature which takes values that belong to all classes. We conclude that the choice of the number of classes is very important. Moreover, every feature used in the training phase should contribute to the prediction of the target class. Our experimental results seem to indicate that a choice of three classes correspond to our need in terms of feature contributions. Fig 4 shows the individual feature contributions for our model after mapping the score value to three classes.

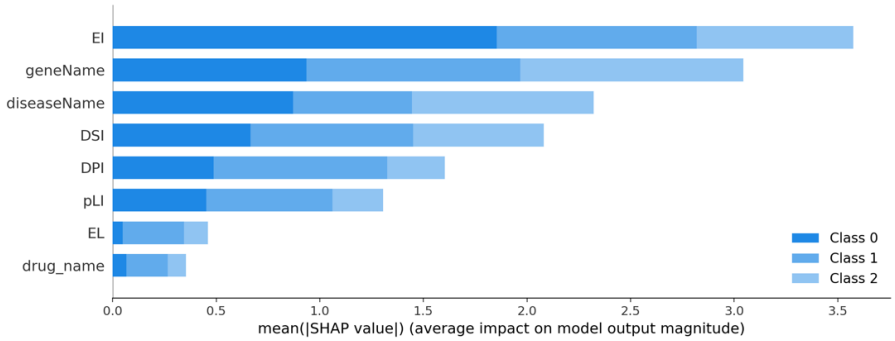

Fig 4. Individual feature contributions based on used sub-dataset, using three classes.

Based on the results we got form our experiments on feature contribution. We manually constructed a new balanced dataset with 10000 samples. Learning from our new balanced dataset, after mapping the score to 3 classes, we successfully increased the performance of our model. We achieved a training score of 0.9877 with a mean cross-validation score of 0.97 and a mean squared error of 0.035. Table 4 shows the results of our trained model, there is a higher performance in term of recall for each class.

Table 4: Results of the 10-fold cross-validation of our balanced dataset.

class	precision	recall	f1-score	count of samples
0	0.99	0.99	0.99	334
1	0.96	0.94	0.95	321
2	0.96	0.97	0.96	345
weighted avg	0.97	0.97	0.97	

We classified the interactions from our graph. Table 5 represents top-20 candidates obtained by our balanced model. We sorted the triplets by descending order of the strength of interactions.

Table 5: List of 20-top candidates of gene-drug-disease triplets predicted from used dataset.

rank	gene	disease	drug
1	EGFR	Non-small cell lung cancer metastatic	Erlotinib
2	NTRK1	Intellectual disability	Suramin
3	ERBB2	Malignant neoplasm of breast	Trastuzumab
4	BRAF	Adenocarcinoma of lung (disorder)	Alpelisib
5	MAP3K1	Non-small cell lung cancer metastatic	Carboplatin
6	KRAS	Adenocarcinoma of lung, stage IV	Atezolizumab
7	DSCAM	Non-small cell lung cancer	Carboplatin
8	KCNJ11	Diabetes mellitus	Naminidil
9	GRB2	Adenocarcinoma of lung (disorder)	Dactinomycin
10	ACE	Congestive heart failure	Cilazapril
11	CDH1	Carcinoma of urinary bladder, superficial	Erlotinib
12	HMGCR	Aicardi-goutieres syndrome 1	Simvastatin
13	GHRL	Diabetes nephropathy	Celecoxib
14	CYP2B6	Adrenal cortical hypofunction	Trofosfamide
15	HLA-DRB1	Endothelial dysfunction	Lym-1
16	BRCA2	Metastatic prostate carcinoma	Talazoparib
17	LEP	Monogenic obesity	Risperidone
18	MET	Malignant neoplasm of kidney	ALTIRATINIB
19	BRCA2	Prostate carcinoma	Evofosfamide
20	ADRB2	Diabetes	Propranolol

In the group of top-20 triplets found with our prediction model (Table 5). It is important to mention that unreleased gene-drug-disease triplets were joined with widely studied candidates in the field of cancerology. As an example, it is commonly recognized that aberrant epidermal growth factor receptor (EGFR) signaling led to varied oncogenic phenotypes [22]. Alongside with our GDD graph, investigations have revealed that the EGFR gene mutation was related with EGFR-

targeted agents' efficacy such as Erlotinib's (rank 1) in the case of non-small cell lung cancer (NSCLC) [27,28].

Contrariwise, the gene MAP3K1 associated with Carboplatin and NSCLC disease (rank 5) appeared novel. On advanced NSCLC patients treated with this antineoplastic chemotherapy drug, the genome-wide association study shows that a single nucleotide polymorphism in the DSCAM gene has been identified as a prognostic biomarker candidate [29]. This sustains our gene-drug-disease triplet in rank 7 and reveals possible MAP3K1-DSCAM interaction which needs to be more studied.

Lung cancer is provoked by the excessive cell development in malignant lung tumor. It is known as the most frequently causes of deaths in men and second in woman. Lung cancer is categorized to two sorts: small cell lung cancer and non-small cell lung cancer. We ranked top-20 predicted candidates related to lung cancer. Table 6 shows the list of gene-drug pairs we have found.

Table 6. List of 20-top candidates of gene-drug pairs predicted from our data related to lung cancer

rank	gene	drug
1	EGFR	Erlotinib
2	BRAF	Alpelisib
3	MAP3K1	Carboplatin
4	KRAS	Atezolizumab
5	DSCAM	Carboplatin
6	GRB2	Dactinomycin
7	CASP8	Conatumumab
8	PIK3CA	Linsitinib
9	TGFB1	Amifostine
10	TNF	Adalimumab
11	ACE	Benazepril
12	TP53	Abemaciclib
13	PIK3CA	Afuresertib
14	PTEN	Abiraterone
15	AKT1	Gigantol
16	HRAS	Trifluoroethanol
17	ELN	Vonapanitase
18	IGF1R	Brigatinib
19	TGFBR2	Galunisertib
20	NTRK3	Radicicol

Among all pairs in the prediction list, there are 18 known causal genes unraveled as true positives. HRAS, the rank 16 gene, belongs to the Ras oncogene family. Malfunctioning in this gene is elaborated in a varied spectrum of cancers. TGFBR2 is a transforming Growth Factor Beta Receptor 2 which may induce esophageal cancer. According to GeneCards database, two genes are susceptive for lung cancer, as for AKT1 and TP53. They contribute in the small cell lung cancer pathway according to PathCards database. Therefore, 18 gene-drug ranked within top-20

have supportive evidence. In addition, all the shown pairs have strong association to linked drugs according to DGIdb.

5 Conclusion

In this work, we proposed a new approach on how to integrate and validate interactions between gene-drug-disease by learning from heterogeneous biological data. We constructed two layers of gene-disease and gene-drug interactions to build an integrated network. Afterwards, we used the XGBoost classifier on a set of 10000 interactions in the training phase. Our prediction model was evaluated using several methods and achieved a f1-score of 0.97. Moreover, we used our classifier to identify and rank 20-top gene-drug-disease interactions. The results were interpreted and compared to the medical literature. We have also extracted a list of top-20 gene-drug pairs related to lung cancer which contained numerous known causal genes unraveled as true positive. Results we obtained with our approach are particularly promising in order to formulate new hypothesis about treatments that might provide multiple advantages.

References

[1] Iorio, Francesco et al. (2013) Transcriptional data: a new gateway to drug repositioning?, *Drug discovery today*, 18(7), 350-357.

[2] Booth, Bruce, and Rodney Zemmel. (2004) Prospects for productivity, Nature reviews. *Drug discovery*, 3(5), 451–6.

[3] Li, Jiao et al. (2016) A survey of current trends in computational drug repositioning, *Briefings in bioinformatics*, 17(1) 2–12.

[4] Institute of Medicine (US) Forum on Drug Discovery, (2009) Development, and Translation. Breakthrough Business Models: Drug Development for Rare and Neglected Diseases and Individualized Therapies: Workshop Summary. *Washington (DC): National Academies Press* (US).

[5] Hamza Hanafi, Badr Dine Rossi Hassani, and M'hamed Aït Kbir. 2019. Biological networks analysis, analytical approaches and use case on protein-protein network interactions, *the 4th International Conference on Smart City Applications (SCA '19),* (pp. 1–5). Association for Computing Machinery.

[6] Hamza Hanafi, Fadoua Rafii, Badr Dine Rossi Hassani, and M'hamed Aït Kbir. 2018. A graph based model for multiple biological data sources integration, *the 3rd International Conference on Smart City Applications (SCA '18).* (pp. 1–5) Association for Computing Machinery.

[7] Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege, Andrew Collins. (2013), Machine learning approaches for the discovery of gene–gene interactions in disease data, *Briefings in Bioinformatics*, 14(2), March 2013, 251–260.

[8] Liu, S., Zhang, J., Xiang, Y., Zhou, W., & Xiang, D. (2019). A Study of Data Pre-processing Techniques for Imbalanced Biomedical Data Classification. ArXiv, abs/1911.00996.

[9] Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. (2008) *Science*.322(5903), 881–8.

[10] Vilar, S., & Hripcsak, G. (2017). The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Briefings in bioinformatics*, 18(4), 670–681.

[11] Vilar, Santiago et al. (2012). Drug-drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association: JAMIA*, 19(6), 1066–1074.

[12] Ferdousi, R., Safdari, R., & Omidi, Y. (2017). Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of biomedical informatics*, 70, 54–64.

[13] Zhang, W., Chen, Y., Li, D., & Yue, X. (2018). Manifold regularized matrix factorization for drug-drug interaction prediction. *Journal of biomedical informatics*, 88, 90–97.

[14] Shtar, G., Rokach, L., & Shapira, B. (2019). Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. *PloS one*, 14(8), e0219796.

[15] Park, K., Kim, D., Ha, S., & Lee, D. (2015). Predicting Pharmacodynamic Drug-Drug Interactions through Signaling Propagation Interference on Protein-Protein Interaction Networks. *PloS one*, 10(10), e0140816.

[16] Dhami, D. S., Kunapuli, G., Das, M., Page, D., & Natarajan, S. (2018). Drug-Drug Interaction Discovery: Kernel Learning from Heterogeneous Similarities. *Smart health (Amsterdam, Netherlands)*, 9(10), 88–100.

[17] Simboost He,T. et al. (2017). Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminform*, 9, 24

[18] Zhu S, Bing J, Min X, Lin C and Zeng X (2018). Prediction of Drug–Gene Interaction by Using Metapath2vec. *Front. Genet.* 9, 248.

[19] Piñero, Janet et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic acids research.* 45, 833–839.

[20] Piñero, Janet et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes, *the journal of biological databases and curation*, bav028.

[21] Bauer-Mehren, Anna et al. (2010). DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks, *Bioinformatics (Oxford, England)*, 26(22), 2924–2926.

[22] Mienye, I.D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review. Procedia Manufacturing,

[23] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20 - 28.

[24] Tianqi Chen and Carlos Guestrin. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). (pp. 785–794). Association for Computing Machinery, New York, NY, USA.

[25] Daniel Berrar. (2019). Cross-Validation, Encyclopedia of Bioinformatics and Computational Biology. (pp. 542-545). Elsevier.

[26] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). (pp. 4768–4777) Curran Associates Inc., Red Hook, NY, USA.

[27] Mayo, Clara et al. (2012). Pharmacogenetics of EGFR in lung cancer: perspectives and clinical applications, *Pharmacogenomics*, 13(7), 789–802.

[28] de Mello, Ramon Andrade et al. (2013). EGFR and KRAS mutations, and ALK fusions: current developments and personalized therapies for patients with advanced non-small-cell lung cancer, *Pharmacogenomics*, 14(14) 1765–1777.

[29] Sato, Yasunori et al. (2011). Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel, *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, 6(1), 132-138.

Notes on contributors

Hamza Hanafi received his engineering degree in computer science on 2017 from the Faculty of Sciences and Technologies, University Abdelmalek Essaâdi, Tangier, Morocco. He is currently a Ph.D. candidate, his research interest's computational biology, bioinformatics and machine learning. He has published several research articles in international conferences of computer science.

Badr Dine Rossi Hassani is a full professor of biology at the Faculty of Sciences and Technologies (FST) of Tangier, Morocco, and PhD managing director at LAFILAB laboratory, his research areas interest many disciplines: Cancer Research, Biotechnology, Bioinformatics. He is a member of scientific committees of many international conferences and journals.

M'hamed Aït Kbir is a full professor at the computer science department of the Faculty of Sciences and Technologies (FST) of Tangier, since 2001, University Abdelmalek Essaâdi, Morocco. As a member of LIST laboratory, since 2007, his research works focus on three main areas: - Computer vision (multimedia flow optimization, multimedia document content watermarking, object recognition, 3D contents indexing and retrieval, 3D reconstruction) - Artificial intelligence (machine learning, Deep learning, Planning and search strategies) - Bioinformatics (Micro-array Data Decision Making, biological data integration, Biological Networks analysis). He is a member of scientific committees of many international conferences and journals. As an expert, he participates in the evaluation of public and private education programs for the ANEAQ and the ministry of higher education and scientific research.