An Efficient Machine Learning Method to Solve Imbalanced Data in Metabolic Disease Prediction

Vânia Cecchini
Life Sciences Research Unit
University of Luxembourg
Esch-sur-Alzette, Luxembourg
vania.cecchini@uni.lu

Thanh-Phuong Nguyen
Megeno, Luxembourg
Life Sciences Research Unit
University of Luxembourg
phuong.nguyen@megeno.com

Thomas Pfau
Life Sciences Researh Unit
University of Luxembourg
Esch-sur-Alzette, Luxembourg
thomas@thomaspfau.de

Sébastien De Landtsheer
Life Sciences Research Unit
University of Luxembourg
Esch-sur-Alzette, Luxembourg
sebastien.delandtsheer@uni.lu

Thomas Sauter
Life Sciences Research Unit
University of Luxembourg
Esch-sur-Alzette, Luxembourg
thomas.sauter@uni.lu

***Abstract*—The increase of obesity, its related diseases and the high incidence of metabolic diseases as a whole, constitute a major public health problem on a global scale. New strategies that allow for the discovery of novel metabolic disease-related genes are necessary to develop new treatments. In this paper, we proposed an efficient method to predict metabolic disease genes, solving the problem of imbalanced data. The method combined protein-protein interactions and miRNA-target interactions to construct integrated networks, whose topological properties can be used as features to train machine learning classifiers. We applied different strategies to optimize imbalanced class. The best model of gradient boosting achieved a significant F1-score of 0.82. When testing the model with non-disease genes, we predicted 549 candidates, out of which 123 were validated indirectly from literature to be related to metabolic diseases. The remaining genes' functions were investigated by gene enrichment analysis, revealing their association with diseases known to co-occur with metabolic diseases, such as cancer and cardiovascular conditions. These results indicated that this method contributed to the identification of novel metabolic disease-related genes.**

Keywords—metabolic disease, protein-protein interaction network, miRNA-target interaction, machine learning, disease gene prediction, imbalanced data.

All materials and codes are available upon request.

I. INTRODUCTION

Metabolic disease (MD) is an inclusive term used to describe a large group of diseases that compromise the normal functioning of metabolism. These disorders are distributed into two different categories: inherited or acquired during life. Inherited MDs, also known as inborn errors of metabolism, are diseases that despite being considered as rare individually, have a high incidence when considered as a whole. Their prevalence is estimated to be about 1 in every 1.000 individuals [1]. Obesity is an acquired MD with a large number of comorbidities and its related diseases include other acquired MDs, such as type 2 diabetes, as well as cancer and heart disease, which can lead to early death [2]–[4]. Due to the increase of MDs on a global scale, there is a demand for solutions that can identify genes involved in the emergence of these disorders, in the hopes that these can be used in the development of new therapeutic strategies.

One of such strategies consists of combining protein-protein interactions (PPIs) and miRNA-target interactions (MTIs). The majority of proteins work in protein complexes and proteins that interact with one another are often found to be involved in the same cellular processes, making the study of protein-protein interactions (PPIs) more relevant, namely in protein function and drug target prediction, as well as in disease research [5]–[7]. Due to their regulatory role in important cellular functions, such as metabolism and gene regulation, the interest in the study of miRNAs involvement in disease as biomarkers has recently increased [8], [9]. Consequently, the study of integrated networks consisting of protein-protein interaction networks (PPINs) and miRNA-target interaction networks (MTIs) is likely to provide insights into MD genes. The choice of these networks is related to the commonly accepted principles in bioinformatical protein function prediction that neighboring proteins are more likely to share functions and that miRNAs that target genes play an important role in determining if a gene is disease-related [10]–[13].

The use of ML in the biomedical field has been rapidly increasing in recent years and examples of its use include disease detection and diagnosis, disease prediction and prognosis, gene-disease relations, protein function prediction, among others [6], [7]. Mordelet *et. al.* developed an algorithm for the identification of disease candidate genes using multitask machine learning from positive and unlabeled instances [14]. Gene functional similarities were used to train machine learning classifiers to predict disease genes [15]. miRNA-disease association was predicted by using a boosting algorithm (XGBoost) [16]. In a nutshell, ML is a powerful and versatile tool that when compared to traditional experimental techniques, has the advantage of being significantly faster in terms of workflow and production of results.

However and despite all its advantages, there is a real problem occurring in many ML tasks, which is imbalanced data [17]–[19]. In biomedical datasets, the positive class is commonly much smaller than the negative class and this imbalance makes algorithms biased towards the majority class, thus affecting their ability to classify and make predictions on new data [20]. Given that there is no universal solution to this problem, the best approach seems to be trying different balancing methods, such as assigning larger penalties to wrong predictions, redistributing class weights or over/undersampling techniques, and select the one that yields the best results for the specific task at hands [20].

978-1-7281-3003-3/19/$31.00 ©2019 IEEE

In this paper, we have proposed an efficient computational approach to predict novel MD-related genes by solving a critical problem of imbalanced data. This method combined PPINs, MTIs and ML classification techniques. We used the Synthetic Minority Over-sampling Technique (SMOTE) [21] to upsample the minority class in our data and trained a gradient boosting classifier (GBC) that achieved an F1-score of 0.82 and predicted 549 MD candidate genes. Interestingly, there are 123 genes of the 549 candidate ones that were inferred to be related to MDs via literature mining. We then performed a gene enrichment analysis that showed significant associations between the remaining candidate genes with MD co-occurring conditions. Our results demonstrated that this approach promisingly predicted MD genes.

II. Materials and Methods

A. Datasets

Two integrated networks were constructed from two PPI databases (DBs). One to extract the features for model training and another to make predictions on. To each one of these networks, the MTI data was combined, as well as a list of known MD genes.

1) *Protein-protein interactions.*

The Biological General Repository for Interaction Databases (BioGRID) [22] was used for training the model. BioGRID is a PPI repository, which contains literature curated data for several organisms. This DB was chosen for the model training not only for the quality of its interactions but also by their quantity. BioGRID is a very large database, thus providing a good training set. Version 3.4.154 for *Homo sapiens* was used.

We curated data from the Human Protein Reference Database (HPRD) [23], release9_041310 to perform predictions. This DB also contains curated PPI interactions extracted from literature but just for human proteins. In opposition to BioGRID, HPRD is a rather small DB but it is a well-known and reliable human PPI database.

2) *miRNA-target interactions.*

The experimentally validated miRNA-target interactions database (miRTarBase) [24], was selected because it contains the largest amount of solely experimentally validated MTIs, collected from relevant literature related to miRNA functional studies. The 6.0 release for *Homo sapiens* was downloaded.

3) *Metabolic disease gene list.*

For the generation of this list, an older version of the MD-related genes present in the Comparative Toxicogenomics Database (CTD) [25] was obtained. This DB aims to understand environmental exposures on human health by combining information on chemical-gene, chemical-disease, and disease-gene association. An updated CTD list of MD-related genes was also extracted to be used as validation for the predicted candidate genes.

B. Workflow

The workflow consisted on six major steps, as shown in Fig. 1.

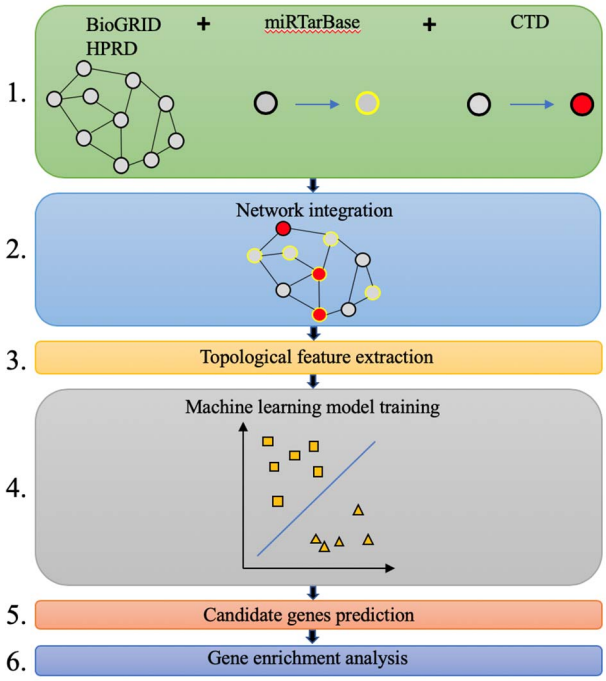


Fig. 1. General workflow containing six major steps

For the data mining step (1), the necessary data for the network integration was gathered to obtain two integrated networks, one for each PPI DB. Using information from three different types of DBs (PPI + MTI + List of known MD genes), the two PPI integrated networks were constructed (2). The next step is to extract topological features associated with each protein present in the integrated network (3). These topological features were the features used to train the machine learning models on BioGRID (4). Once the best predictive model was selected, predictions were made on HPRD, to obtain a list of candidate MD genes (5), followed by a gene enrichment analysis (6).

C. Network integration and feature extraction

For the network integration step, an algorithm adapted from [17] was used. A visual guide of the network integration is presented in Fig. 2.

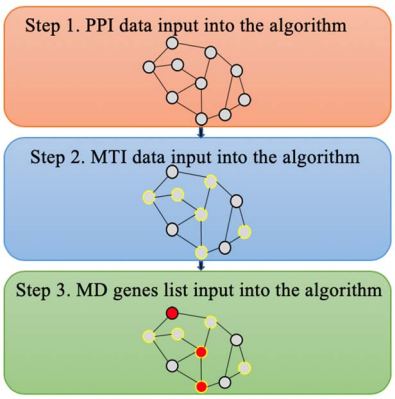
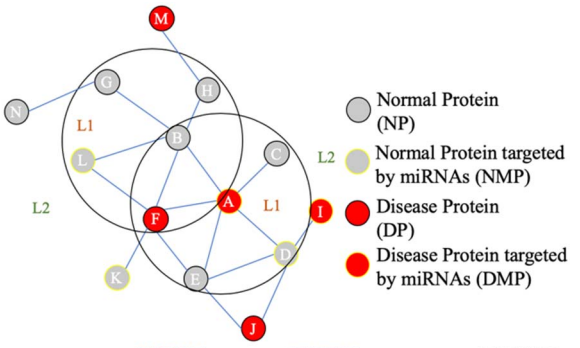


Fig. 2. Network integration schematic consisting of three steps.

The three-step integration was performed for both DBs, and the process consisted of first uploading and then integrating the PPI and MTI networks, resulting in an intermediate network consisting of two types of nodes: normal proteins (NPs) and normal proteins targeted by miRNAs (NMPs). To this intermediate network, a list of known MD

genes was added, producing a final network with four types of nodes: NPs, NMPs, disease proteins (DPs) and disease proteins that are targeted by miRNAs (DMPs). After the integration, the radius of each network was computed to find out the maximal length of the shortest network. This step was performed to avoid including in the training data, features which might not be available when forming new predictions, like information about very distant nodes. The output files from the network integration consist of protein topological vectors that served as the features from which the ML algorithms were trained. A toy model of both length and vectors is presented in Fig. 3.



		LENGTH 1				LENGTH 2					LENGTH Max			
PROTEIN	**CLASS**	NP 1	NMP 1	DP 1	DMP 1	NP 2	NMP 2	DP 2	DMP 2	...	NP Max	NMP Max	DP Max	DMP Max
A	MD	3	1	1	0	5	3	2	1	...	...	...	...	...
B	nonMD	2	1	1	1	5	3	2	1	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 3. Extracted features for machine learning model training through different network lengths.

The topological features consist of four different types of nodes showing the number of proteins, which belong to a specific type of node that are at a maximal shortest path distance from the protein of interest. This maximal distance varies between the shortest length, which is one, and the maximal network length. In the above example, protein A is the protein of interest. For length 1, i.e. the immediate first neighbor, the topological vector is as follows: A = (3,1,1,0). This means that at a maximal distance 1, protein A interacts with 3 NPs, 1 NMP, 1DP and 0 DMPs. If we now consider length 2, vector A will be: A = (3,1,1,0,5,3,2,1). This vector will increase until the maximal distance of length Max, or the maximal length of the network.

D. *Imbalanced data*

SMOTE [21] is an upsampling technique that uses k nearest neighbors to create new synthetic minority instances based on the real ones. The sampling ratio can go until one, meaning that the classes are fully balanced. However, to avoid the possible leakage of new data points into the test set and consequent overfit and misleading accuracy scores, SMOTE should only be used in the training set, after the train/test split.

E. *Machine learning*

For our binary classification problem, a boosting algorithm (GBC) was trained. Boosting consists in training an ensemble of weak learners in the form of decision trees, with the goal of creating a final strong learner. In this iterative method, misclassified samples, i.e., samples that are hard to classify, gain weight and allow the weak learners to learn from them, thus improving the ensemble performance [20].

III. RESULTS AND DISCUSSIONS

Because no wet-lab experimental validation was made on the predicted candidate genes, an older version of the known MD genes list was used to build the PPINs with the purpose of seeing if any of the predicted candidate genes is present in the updated MD gene list extracted from CTD, serving as a form of validation. The remaining candidate genes not validated as MD genes, were the target of a gene enrichment analysis, using gene enrichment tool Enrichr [26]. Enrichr is a comprehensive resource, containing 184 annotated gene sets from 102 gene set libraries.

All the computational framework was performed using Jupyter lab running Python 3.7.2.

A. *Data statistics pre-machine learning.*

After the data curation, the following statistics were obtained and are displayed in Tables I and II.

TABLE I. Total number of metabolic disease genes extracted from the CTD database

	MD genes
# genes	1368
# genes targeted by miRNAs	1020

TABLE II. Data extracted from both PPINs and statistics including metabolic disease genes and miRNA-target information

	BioGRID	**HPRD**
# interactions	401710	39046
# genes	18233	9455
# MD genes	1159	942
# nonMD genes	17074	8513
# MD genes targeted by miRNAs	13220	816
# nonMD genes targeted by miRNAs	982	7028

A tendency for a high frequency of miRNA-target interaction towards MD genes seems to be apparent when looking at tables I and II, supporting the idea that miRNAs do play an important role in determining whether a gene is related to disease.

B. *Predicting MD disease genes using gradient boosting classifier*

1) Imbalanced data. The BioGRID training data (Table II), is quite imbalanced (1159 MD genes to 17074 nonMD genes). To approach this problem, a stratified train/test split was performed on the training set with a ratio of 80/20 and a for loop was computed to test several balancing SMOTE ratios with the GBC, to retrieve the best performing ratio. The accuracy scores decreased as the upsampling ratio in the training set increased, ranging from the accuracy value of 0.93 for a balancing ratio of 0.2, to 0.79 for a balancing ratio of 1. In the imbalanced test set, the predictor classified the majority of instances as belonging the nonMD class, and because the MD class is much smaller than the nonMD, the accuracy score is high and therefore misleading. As the ratio increases, so does the number of correctly predicted MD instances but also do the nonMD instances that are predicted as MD (false positives), thus decreasing the accuracy score.

For the final model training, the SMOTE ratio of 1 was used and the selected measure to evaluate model performance was the F1-Score. The number of training and test set instances are shown in Table III.

TABLE III. MD and nonMD instances in the training set after SMOTE upsampling on the MD class and test set instances

	MD	**nonMD**	**Total**
Training set	13659	13659	27318
Test set	232	3415	3647

2) Feature importance. To understand which features are the most important ones for the classifier to decide to which class each sample belongs to, this information was extracted (Fig. 4).

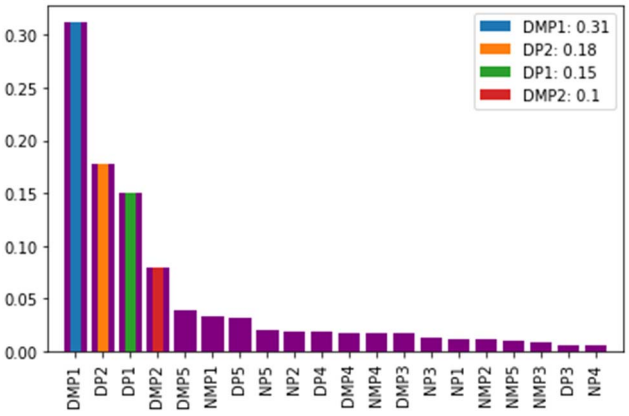


Fig. 4. Feature importance for the Gradient Boosting Classifier

The four most relevant features for the classifier are DMP1 with an importance of 0.31, DP2 with 0.18, DP1 with 0.15, and finally DMP2 with 0.1. All features correspond to disease proteins, two of them targeted by miRNAs (DMP1 and DMP2). Two of them are the first immediate neighbors (DMP1 and DP1), and the other two the second immediate neighbors (DP2 and DMP2) of the proteins of interest. These results support both the assumptions on which this framework was built on. The first one is that neighboring proteins are more likely to share functions and the second one that proteins that are targeted by miRNAs play an important role in determining whether a protein is related to disease. This is especially true for the most important feature DMP1: a disease protein, targeted by miRNAs that is in the immediate vicinity of the protein that will be classified by our model.

3) Model training and evaluation. A GBC was trained using the balanced data with a stratified 10-fold cross-validation. The model was then tested on the imbalanced test set and the evaluation metrics are presented in Table IV.

TABLE IV. Evaluation metrics obtained for the test set

Accuracy	0.77
Precision	0.91
Recall	0.77
F1-Score	0.82
AUC	0.72

The accuracy scores obtained for the training set (0.79) and test set (0.77) were very similar, meaning that the model is not overfitting and is a good predictor for new unseen data.

For the model not to randomly classify the samples as being positive, the number of TPs should be high, the number of FPs low and the AUC (area under the ROC curve) score higher than 0.5. The obtained ROC curve with an AUC of 0.72 is presented in Fig. 5.

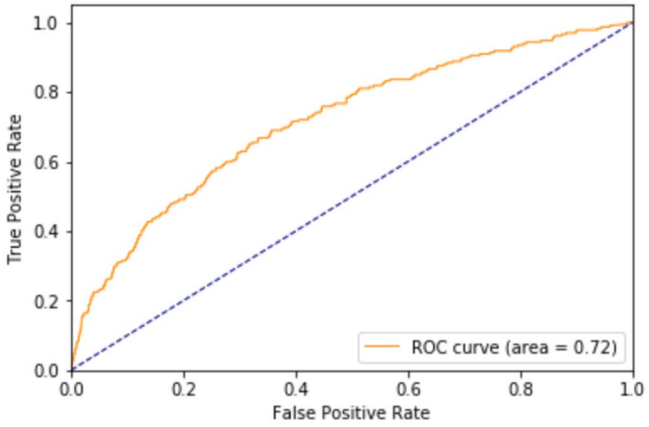


Fig. 5. ROC curve obtained from the GBC.

To better evaluate the model' performance, regarding the goal of this paper, we looked at the precision (0.91), recall (0.77), and their harmonic mean, the FI-score (0.82).

C. Predictions

To test the performance of the model, we used HRPD data for predicting new candidate genes (Fig. 6).

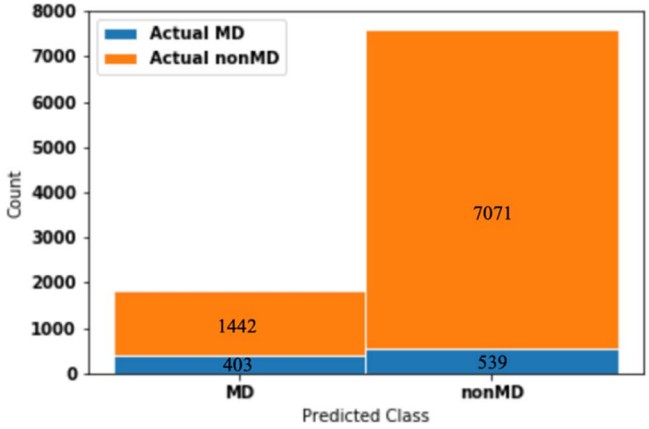


Fig. 6. MD and nonMD predicted genes for the HPRD DB

HPRD is also highly imbalanced, 942 MD to 8,513 nonMD. The classifier correctly predicted less than half the actual MD instances (403) and in addition to the imbalance, these results can be also explained by the difference between both network properties. We predicted 1,442 new candidate genes and to reduce this number, the median of the obtained prediction probabilities was computed (0.67) and a cut-off was performed, reducing the initial list of candidate genes from 1,442 to 549.

D. Validation

The CTD list used for the PPIN integration has 1,368 MD genes and the updated CTD list used for validation 3,183. The final list of 549 predicted candidate genes was compared to the updated CTD list of known MD genes to look for

overlaps. A total of 123 predicted candidate genes was confirmed as being a known MD gene, supporting the idea that the proposed framework can help improve the identification of novel MD genes. The hypergeometric p-value of the validated 123 genes is 2.6645e-15, showing the statistical significance of our results. The remaining 426 candidate genes were subjected to an enrichment analysis performed on the Enrichr [26] tool. The obtained results show the associations between the candidate genes and diseases obtained in the Online Mendelian Inheritance in Man (OMIM) [27] DB (Fig. 7 - A) and PPINs connecting disease genes from OMIM (Fig 7 - B).

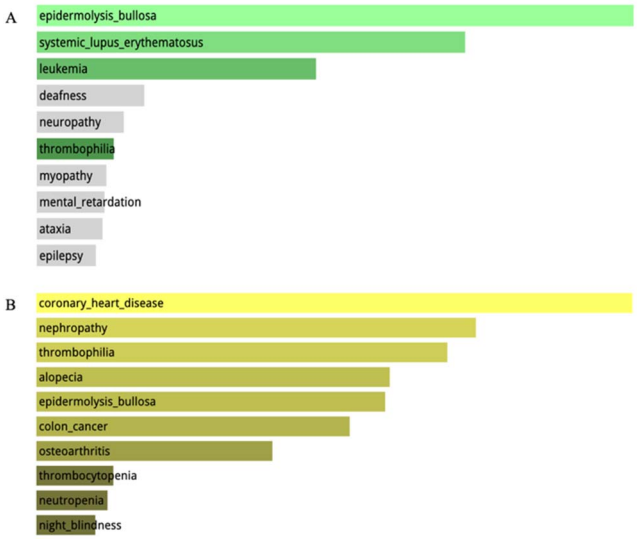


Fig. 7. Gene enrichment analysis performed using Enrichr, showing the results obtained for the OMIM DB.

This analysis revealed the association of these candidate genes with autoimmune diseases like systemic lupus erythematosus, cardiovascular problems, such as myocardial infarction and coronary heart disease, as well as leukemia and colorectal cancers. Given the comorbidity that these diseases have with MDs, these genes seem to be suitable MD candidate genes.

IV. Conclusion

In this paper, we presented a pipeline that can contribute to the prediction of disease candidate genes by combining PPIN properties, MTIs, data balancing techniques and ML classification. From the predicted candidate genes, 123 genes were validated as MD genes and the remaining ones were associated with diseases known to co-occur with MDs, therefore reinforcing the importance of PPINs and miRNAs in novel disease candidate genes prediction, as well as the importance of dealing with the imbalanced data problem in ML. These achieved results showed the potential in the identification of novel disease candidate genes.

V. References

[1] M. Crook, "Atlas of Metabolic Diseases.," *J. Clin. Pathol.*, vol. 53, p. 947, 2000.

[2] J. S. Garrow, "Obesity and related diseases.," *Obes. Relat. Dis.*, 1988.

[3] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The Disease Burden Associated With Overweight and Obesity," *JAMA*, vol. 282, no. 16, p. 1523, Oct. 1999.

[4] T. Kadomatsu, M. Tabata, and Y. Oike, "Angiopoietin-like proteins: emerging targets for treatment of obesity and related metabolic diseases," *FEBS J.*, vol. 278, no. 4, pp. 559–564, Feb. 2011.

[5] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Protein-protein interaction detection: methods and analysis.," *Int. J. Proteomics*, vol. 2014, p. 147648, Feb. 2014.

[6] J. De Las Rivas and C. Fontanillo, "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.," *PLoS Comput. Biol.*, vol. 6, no. 6, p. e1000807, Jun. 2010.

[7] F. Jordán, T.-P. Nguyen, and W. Liu, "Studying protein-protein interaction networks: a systems view on disease," *Brief. Funct. Genomics*, vol. II, no. 6, pp. 497–504, 2012.

[8] A. M. Ardekani and M. M. Naeini, "The Role of MicroRNAs in Human Diseases.," *Avicenna J. Med. Biotechnol.*, vol. 2, no. 4, pp. 161–79, Oct. 2010.

[9] D. Pallez, J. Gardès, and C. Pasquier, "Prediction of miRNA-disease Associations using an Evolutionary Tuned Latent Semantic Analysis," *Sci. Rep.*, vol. 7, no. 1, p. 10548, Dec. 2017.

[10] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, Jul. 2006.

[11] X. Chi and J. Hou, "An iterative approach of protein function prediction," *BMC Bioinformatics*, vol. 12, no. 1, p. 437, Nov. 2011.

[12] S. Moosavi, M. Rahgozar, and A. Rahimi, "Protein function prediction using neighbor relativity in protein–protein interaction network," *Comput. Biol. Chem.*, vol. 43, pp. 11–16, Apr. 2013.

[13] D. D. Truong, C. S. N. Ngoc, V. T. Nguyen, M. T. Tran, and A. D. Duong, "Knowledge and Systems Engineering," *Adv. Intell. Syst. Comput.*, vol. 244, pp. 401–413, 2014.

[14] F. Mordelet and J.-P. Vert, "ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples.," *BMC Bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[15] M. Asif, H. F. M. C. M. Martiniano, A. M. Vicente, and F. M. Couto, "Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology," *PLoS One*, vol. 13, no. 12, p. e0208626, Dec. 2018.

[16] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction," *Cell Death Dis.*, vol. 9, no. 1, p. 3, Jan. 2018.

[17] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies *," 2000.

[18] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 2009, pp. 875–886.

[19] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[20] S. Raschka and V. Mirjalili, *Python Machine Learning, 2nd Edition*, 2nd ed. Packt Publishing Ltd, 2017.

[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[22] A. Chatr-aryamontri *et al.*, "The BioGRID interaction database: 2017 update," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D369–D379, Jan. 2017.

[23] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database--2009 update," *Nucleic Acids Res.*, vol. 37, no. Database, pp. D767–D772, Jan. 2009.

[24] C.-H. Chou *et al.*, "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D239–D247, Jan. 2016.

[25] A. P. Davis *et al.*, "The Comparative Toxicogenomics Database: update 2017.," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D972–D978, Jan. 2017.

[26] M. V. Kuleshov *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W90–W97, Jul. 2016.

[27] McKusick-Nathans Institute of Genetic Medicine, M. and N. C. for B. I. Johns Hopkins University (Baltimore, and M. National Library of Medicine (Bethesda, "OMIM - Online Mendelian Inheritance in Man," 2017. [Online]. Available: https://www.omim.org/. [Accessed: 27-Dec-2017].