IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117113

Prediction of Mechanism of Action (MoA) of Novel Drugs

Akshay A Kumar[1], Anurag Ashish Khot[2], and Merin Meleet[3]

R.V. College of Engineering, Bengaluru, Karnataka, 560059[1,2,3]

Abstract The term Mechanism of Action (MoA) is a biochemical interaction by which the drug produces its effect on the cell. A drug's precise molecular targets, such an enzyme or receptor, to which it binds are often included in a mechanism of action. A live creature being exposed to a chemical results in functional or structural changes at the cellular level that are referred to as a mode of action (MoA). Mechanism of Action of a drug makes a drug more usable. If the MoA of a drug is known it can be used in various different situations and in proper dosage, the risk of allergies and problems in specific people also reduces. So, prediction of Mechanism of Action of a drug opens up new possibilities for the drug and makes it safer to use.Conducting the following study by performing traditional ANN approach we can conclude that Artificial Neural Networks (ANN) have the capability to predict the mechanism of action, and perform a multilabel classification for a multifeature drug profile, that includes cell viability and gene expression data. feature engineering and tuning the ANN model will yield better results.

Keywords: ANN,Mechanism of Action, gene expression, cell viability, drug

1 INTRODUCTION

The term Mechanism of Action (MoA) is a biochemical interaction by which the drug produces its effect on the cell. A drug's precise molecular targets, such an enzyme or receptor, to which it binds are often included in a mechanism of action. A live creature being exposed to a chemical results in functional or structural changes at the cellular level that are referred to as a mode of action (MoA). Traditional Artificial Neural Network methods were conducted. Perform data preprocessing and data cleaning techniques such as handling duplicate entries, converting data types. Perform Exploratory Data Analysis, the different variables of the dataset and plotting their graph to understand the data better identify patterns to help us make necessary tuning. Use different techniques. Do some performance optimization, hyper-parameter tuning to get better results. Use different performance metrics for multi-label classifiers to conclude which model gives the best results.

the dataset that combines information on gene expression and cell viability. The information is based on a novel technique that concurrently assesses (within the same samples) the pharmacological responses on human cells of 100 distinct cell types, (thus solving the problem of identifying ex-ante, which cell types are better suited for a given drug).Features for the dataset included, gene expression data, and cell viability data. a feature that is indicated as samples treated with a compound or with a control perturbation; control perturbations have no MoAs; time that has 3 values 2, 28 or 72 hours and dose has 2 values high and low.

The data set was obtained from the Broadway Institute Connectivity Map project. For implementation Python 3.7, programming language. Google Collab was used to write scripts and make use of the installed libraries in the cloud for documentation displaying the result. Used other open source python libraries for Data Visualization and Machine Learning like: Numpy, Scikit-learn, Pandas, Matplotlib, NetworkX, Keras and Tensorflow.

2 LITERATURE REVIEW

In [1] the authors studied and compared the classic ensemble-based tree classifier which was trained on mophorphical features which are experimentally extracted in lab environment and Convolutional Neural Network, the other model on comparison which was trained on images itself across distinct cells. As a result, our findings reveal that an ensemble-based tree classifier and a CNN classifier exhibit about equal performance when trained and forecasted on the same cell line in compound MoA prediction for the majority of cells. The ensemble-based tree classifier experienced a loss in classification accuracy, while the CNN classifier saw a fall in prediction performance when the classifiers were applied to unknown cell lines. However this study was concerned with the generalizability of the classifiers across cell lines, so less focus on predictive performance.Not enough Image augmentation was done. evaluated their approach on a small number of drugs.

In [2] The major goal was to concentrate on cutting-edge methods that haven't gotten enough attention but have a lot of

DOI: 10.17148/IJARCCE.2022.117113

promise to progress the industry. The methods of reinforcement learning (RL), transfer learning, and multitask learning are discussed. The results showed that examples and MLT approaches might get around problems with traditional methodologies. When there isn't enough large data, use transfer learning and multitask learning, plus BNN to prevent overfitting. The development of new drugs may entail emerging methodologies. Although other approaches that appeared to have more promise in the larger area of drug development were taken into consideration, no comparative analysis was provided, and no model was thoroughly examined to determine its full potential.

In [3] the objective was to potentially raise awareness of sex differences during clinical decision making. To build a pharmacovigilance algorithm that leverages advances in machine learning to predict sex risks. The results indicated that of the 8.8 million patients in FAERS, 61.9% were female. Sex-differential drug exposure was evident from the biased reporting of sex for certain medications.No significant association was found between concomitant drug exposures and sex. reduction in classification accuracy, the CNN classifier saw a decrease in prediction performance.However it does not work on different forms of data available and machine learning techniques, or other factors apart from sex, including age.

In [4] the work carried out was to provide a machine learning model to predict the mechanism of action of a drug using, Binary Relevance K Nearest Neighbors, Multi-label K-Nearest Neighbors and a custom neural network. These models are evaluated using the mean column-wise log loss.The results were that the custom neural network model had the best accuracy with a log loss. For BRKNN the score improves from 3 neighbors to 10 neighbors. Afterwards, scores decline as the number of neighbors increases. For MLKNN the score improves from 3 neighbors to 20 neighbors. However the models in the prepared were not easily able to accommodate new sources of information as they become available.

In [5] A drug-target prediction platform was developed by Neel S. Madhukar et al. that employs a Bayesian technique to integrate a variety of disparate data kinds impartially and offers a platform that enables the straightforward integration of additional data types as they become available. The outcome was an integrated strategy that improves accuracy. The programme successfully anticipates certain target interactions. Novel microtubule-targeting drug discovery reveals the selective opposition between ONC201 and DRD2 can pinpoint certain pharmacological mechanisms. However, this study makes no conceptual advances. It has not been compared to recent studies in the same field that use other methods that can handle new data types as they become available.

In [6] Giovanna Maria Dimitri et. al. Created clusters of drugs according to several various profiles Discover interaction between groups of drugs sharing similar chemical and protein interaction profiles, side effects and pathways. They came up with a new machine learning tool for drug side effects prediction: DrugClust was presented. The results showed the validity of their approach, obtaining better results in most of the cases. However, this was done only for a certain type of drugs, so not much generality of samples was there. Parameters related to the human body/cell were not considered while checking the side effects of the drug.

in [7] In order to forecast the CGPIs of medications, Chengyou Liu, Andrew M. Hogan, Hunter Sturm, Mohd Wasif Khan, Md. Mohaiminul Islam, A. S. M. Zisanur Rahman, and Rebecca Davis trained GCN-based algorithms. to utilize a large-scale chemical genetic dataset to test a library of 47,272 chemicals using hundreds of mutant cultures of the bacteria Mycobacterium TB that were largely deprived of key proteins. The study of gene clusters compared the D-MPNN with the RDKit descriptor's performance to that of the five baseline models. No matter how many extra molecular descriptors were included, the D-MPNN classifiers still performed best overall in terms of AUROC. However, using solely M. tuberculosis CGIPs to predict MOA.

In [8] Application two classes of methods are done, sequential learning and recommender systems, to the field of pharmaceutics and drug discovery, drug testing. There are hundreds of online publicly curated databases which, provided some scripting efforts, can be integrated modularly to drug development pipelines. This data can be used on ML models to find new observations. Refined DL architectures and sequential algorithms can be used in drug discovery, and drug testing. However, no actual implementation is done to predict the validity of the algorithms considered to verify their effectiveness.Only general cases are considered, specific cases for the drugs and cells are not considered. Which model might suit for what type of drug or cell is not evaluated.

In [9] the objective was to do image-based profiling using machine and deep learning, to improve extraction of relevant signals from profiles. ML and deep learning techniques gave better results than compared to classical approaches to image profiling. No particular gap as it is a review paper. Further investigation of the use of Ml and deep learning in image profiling can be done.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117113

3 HYPOTHESIS

There are works done on different kinds of data trying to predict the mechanism of action of compounds, works done on a narrow subset of them, or directly on the images, exhibiting phenotypic changes on reaction with the compounds and different forms of data.

After conducting the literature survey of all the work done the area of predicting mechanism of action of drugs, the major gaps that we came by was that, there was more comprehensive data available that has more number of instances, higher precision and greater number of features, greater number of features indicate the depth to which the the compounds or drugs in our case, are profiled, since we have data that has been explicitly profiled already, it reduces the work of our model, where in some of the previous works the model had to determine first the effect of the compound on the cell through changes observed in the phenotypes by reading through the images and then determine the mechanism of action using the information. The primary objective of the study was to test the capability of the ANN in predicting the mechanism of actions as a multilabel classifier drug profile data. The drug profile consists of numericals values such as gene expression and cell viability data that is a measure of the expression of genes in a cell and the percentage of living to dead cells respectively on reaction with the compound or drug.

Its hypothesized that Artificial Neural Networks can classify and predict labels, Mechanism of Action for drugs and consequently aid in the field of discovering novel drugs. The work done by Broadway Institute, creating and analyzing large perturbational datasets, this dataset is cleaned, analyzed and preprocessed and provided us with the right data to tackle the objective.

4 METHODOLOGY

The domain being fairly new, the features had to be understood even outside of the machine learning context. The data also looked huge enough with 876 features out of which 772 gene expression data and 100 cell viability data. Testing the capability of ANN meant we required a standard value to compare to, for which we built a baseline ensemble one vs all for 206 classes of Mechanism of actions. Traditional and standard ANN/DNN steps were implemented, feature engineering and hyper parameter tuning was done, the methodology consisted of 4 major steps, which are as follows:

1. Perform data preprocessing and data cleaning techniques such as handling duplicate entries, converting data types. Perform Exploratory Data Analysis, the different variables of the dataset and plotting their graphs and correlations to understand the data better identify patterns, principality and redundancy of the features of the data. It is important to understand the output labels as well, since they are 206 in number, there could arise correlations using co-occurrence networks within the classes that would to understand the nature of the data and problem better and to help make necessary tuning.

2. Since there are a huge number of features, and inferences drawn from the correlation matrices we can choose to perform dimensionality reduction. Results from correlations within the features themselves and with the target, moving variance is plotted with instances of the data. Training loss can be plotted for different numbers of principal components and checked when there is loss of data and accuracy, to determine the number of dimensions that can be reduced. To understand if there was any correlation or co-occurrences in occurrences of MoA, a spring plot can be plotted and clustering of the instances can be done, if there is a significant relationship among them, new features could be prepared, using that clustered label that could higher degree of information.

3. ANN are dealt with many hyper parameters apart from the weights of the nodes, parameters like activation function in each layer of the network, batch size for stochastic gradient descent, choosing of the ideal learning rate, number of nodes and layers and epochs so as to prevent overfitting, underfitting, overshooting, prevent randomization, so model should not be to complex for simpler data comparatively, factors like resulted in preparing multiple models with different parameters to understand which combination gave better results in terms of cross validation accuracy and loss, this cross validation set included unseen data by the model, which is a equivalent representation of the training data in terms of distribution and randomness included 20% of the overall data.

4. After choosing the best overall model with the right parameters, to make use of the complete data this model is to be trained with 5-fold cross validation set, run for a longer number of epochs and smaller learning rate to get to least error, logloss metric was used as the error while training, This will be the final training of the model. The final accuracy and loss will be calculated with yet another set of unseen data which will be used to evaluate the performance of the model compared to other models and the standard ensemble model. The predictions of different MoA can be plotted to

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117113

get birds' view of the predicted output and can be checked to see any anomaly with certain MoAs that can then be dealt accordingly.

Taking care of methodologies is very important so is having clarity on it. Designing is the basic requirement for any kind of project development. Hence, having a detailed plan helps in the smooth development of the project.

5 RESULT AND ANALYSIS

This project uses ANN for implementing the multi-label classification. Model with 1 hidden layer was used as the baseline model, which had 256 nodes in the hidden layer and used sigmoid as the activation function in the output layer. It was trained for 35 epochs. Validation loss was: 0.01572 and validation accuracy was: 0.1181. Training loss was: 0.0146 and training accuracy was: 0.1608.

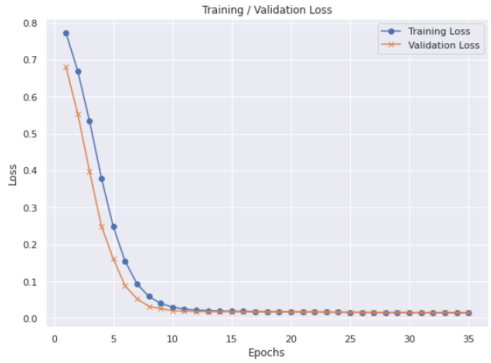

Fig 1: Training and validation losses of the model with 1 hidden layer over the epochs

Fig 2: Training and validation accuracies of the model with 1 hidden layer over the epochs.

As good accuracy was obtained, this was kept as a baseline model and further changes to ANN were tried. Next, Model with 2 hidden layers was tested. It had 256 nodes in both hidden layers. The performance was similar to the model with 1 hidden layer. The coefficient matrix was built and it was found that there is a lot of correlation between the few features. So, PCA was done. The number of gene expressions were 772 and were reduced by a factor of 100. Till 463 features the performance remained the same after that losses increased and accuracy decreased. For cell viability features there were 100 features initially which were reduced by a factor of 10. Till 60 features the performance remained similar after that it deteriorated. So, after PCA the total features became 563, out of which 463 were gene expression data and the other 100 were cell viability data. This data was then fed to the Model with 3 hidden layers. This gave a better performance. After this Hyperparameter tuning was done. In this the model was run multiple times for different hyperparameters. It was found that '**softmax**' as activation function in the output layer performed better than '**sigmoid**' function. The number of nodes in the hidden layers were 512, 256 and 256. In the output layer there were 206 nodes for all the different MoA's. To prevent overfitting the optimal number of epochs were found to be 60. The optimal batch size was found to be 1024. This model with 3 hidden layers was the one which performed the best.

Since, so many models were trained a number of times. A Learning rate function was developed to find the optimal learning rate so that the model training becomes efficient and faster.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117113

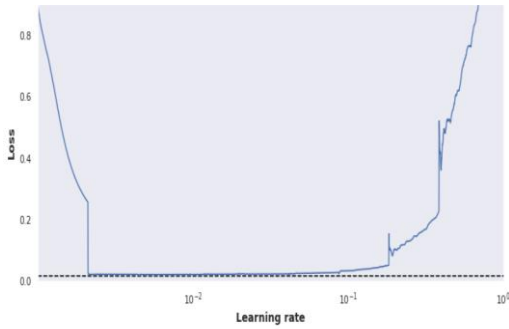

Fig 3: Loss vs Learning rate for model with 3 hidden layers.

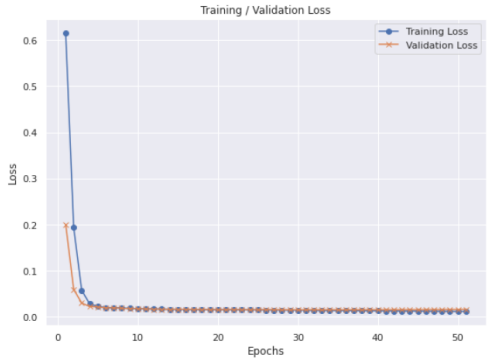

Fig 4: Training and validation losses of the model with 3 hidden layers over the epochs.

Fig 5: Training and validation accuracies of the model with 3 hidden layers over the epochs.

The validation loss for the model with 3 hidden layers was: 0.01548 and validation accuracy was: 0.1205. Training loss was: 0.0118 and training accuracy was: 0.2091.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117113

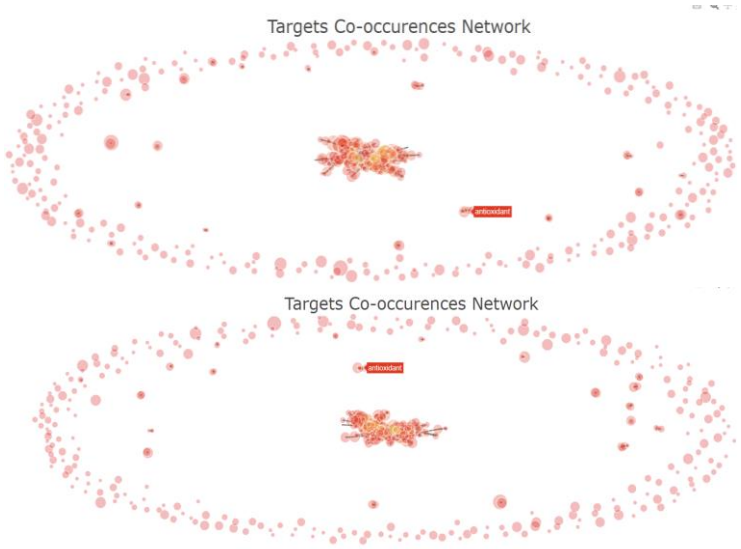

Fig 6: Top: Co-occurrences network of MoA's on training data.
Bottom: Predicted co-occurrences network of MoA's on test data.

In fig 6, the top image shows the co-occurrences network of MoA's for training data. The structure of the co-occurrences network of test data resembles the structure of the co-occurrences network of training data. There is a large cluster at the center in both and some smaller clusters of similar structure as well. Other unrelated MoA's form the circumference in both the cases. For instance, antioxidant in the highlighted smaller clusters occurs with capillary stabilizing agent, nitric oxide scavenger etc. in both the co-occurrences network which shows that similar MoA's are clustered together in both the cases.

6 CONCLUSION AND FUTURE WORK

Prediction of MoA of novel drugs is a very important task. It improves the usage and trust on the drug. It also makes it easier for doctors to suggest and get to know what areas a drug can affect or what can be its side effects. So, prediction of MoA is gaining a lot of scope in the research. In this project it was checked the capability of Artificial Neural Network for the prediction of the MoA.The model with 3 hidden layers gave the best accuracies and the least losses.If there is more vertical data i.e more number of drugs that are tested and more horizontal data i.e more features for the cells being tested then the predictions and accuracies both can improve. There is a lot of future scope for ML in predicting MoA's as other physical and chemical methods for predicting the MoA's are becoming obsolete and as the number of novel drugs being invented is increasing the requirement for better and faster methods like ML is increasing.

REFERENCES

1. Warchal SJ, Dawson JC, Carragher NO. Evaluation of Machine Learning Classifiers to Predict Compound Mechanism of Action When Transferred across Distinct Cell Lines. SLAS DISCOVERY: Advancing the Science of Drug Discovery. 2019;24(3):224-233. doi:10.1177/2472555218820805
2. Moe Elbadawi, Simon Gaisford, Abdul W. Basit, Advanced machine-learning techniques in drug discovery, Drug Discovery Today, Volume 26, Issue 3, 2021, Pages 769-777, ISSN 1359-6446, https://doi.org/10.1016/j.drudis.2020.12.003
3. Payal Chandak, Nicholas P. Tatonetti, Using Machine Learning to Identify Adverse Drug Effects Posing Increased Risk to Women, Patterns, Volume 1, Issue 7, 2020, 100108, ISSN 2666-3899, https://doi.org/10.1016/j.patter.2020.100108.
4. Gururaj, H.L., Flammini, F., Kumari, H.A.C. et al. Classification of drugs based on mechanism of action using machine learning techniques. Discov Artif Intell 1, 13 (2021). https://doi.org/10.1007/s44163-021-00012-2
5. Madhukar, N.S., Khade, P.K., Huang, L. et al. A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun 10, 5221 (2019). https://doi.org/10.1038/s41467-019-12928-6
6. Dimitri GM, Lió P. DrugClust: A machine learning approach for drugs side effects prediction. Comput Biol Chem. 2017 Jun;68:204-210. doi: 10.1016/j.compbiolchem.2017.03.008. Epub 2017 Mar 30. PMID: 28391063.
7. Liu, C., Hogan, A.M., Sturm, H. et al. Deep learning-driven prediction of drug mechanism of action from large-scale chemical-genetic interaction profiles. J Cheminform 14, 12 (2022). https://doi.org/10.1186/s13321-022-00596-6

DOI: 10.17148/IJARCCE.2022.117113

8. Réda C, Kaufmann E, Delahaye-Duriez A. Machine learning applications in drug development. Comput Struct Biotechnol J. 2019 Dec 26;18:241-252. doi: 10.1016/j.csbj.2019.12.006. PMID: 33489002; PMCID: PMC7790737.
9. Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D. et al. Image-based profiling for drug discovery: due for a machine-learning upgrade?. Nat Rev Drug Discov 20, 145–159 (2021). https://doi.org/10.1038/s41573-020-00117-w
10. Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, Hobson S, Mathieson T, Perrin J, Raida M, Rau C, Reader V, Sweetman G, Bauer A, Bouwmeester T, Hopf C, Kruse U, Neubauer G, Ramsden N, Rick J, Kuster B, Drewes G. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol. 2007 Sep;25(9):1035-44. doi: 10.1038/nbt1328. Epub 2007 Aug 26. PMID: 17721511.
11. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011 Oct 30;29(11):1046-51. doi: 10.1038/nbt.1990. PMID: 22037378.
12. Connectivity Mapping: Methods and Applications Alexandra B. Keenan, Megan L. Wojciechowicz, Zichen Wang, Kathleen M. Jagodnik, Sherry L. Jenkins, Alexander Lachmann, and Avi Ma'ayan Annual Review of Biomedical Data Science 2019 2:1, 69-92

