ISSN No:-2456-2165

Deep Action: An approach on the basis of Deep Learning for the Prediction of Novel Drug-Target Interactions

Hilma K K., Karthika Krishnan, Lima P Subran
Department of Computer Science and Engineering
Mar Athanasius College of Engineering
Ernakulam, India

Prof. Neetha Joseph
Faculty in charge, Department of Computer Science and Engineering Mar Athanasius College of Engineering Ernakulam, India

Abstract:- In the processes of drug development and discovery, Drug-target interactions (DTIs) take part a vital position. DTI prediction through laboratory experiments consumes a lot of time. Also they were costly and tiring. Although computational approaches can recognize new interactions between drug-target pairs and speed up the drug conversion procedures, some problems like large scope of data and imbalanced class have been encountered in the course of the prediction procedures, and the number of unknown interactions were huge. Therefore, an approach on the grounds of deep learning (deepACTION) is put forward to predict possible or unrevealed DTIs. Here, each drug chemical structure and protein sequence is transformed according to structural and sequence information using different descriptors to correctly constitute their properties. In this method the majority and minority instances in the dataset are balanced using the SMOTE technique. For accurate DTI prediction a convolutional neural network (CNN) algorithm is trained with balanced and reduced features. For comparing the performance of the DeepACTION model with that of other methods AUC is regarded as the primary evaluation metric. An AUC curve of 0.933 is achieved by Deep ACTION model for the experimental dataset acquired from the Drug Bank database. Based on exper-imental results it is evident that the model is capable to predict a remarkable number of new DTI's and it produce thorough knowledge that inspires scientists to instigate advanced drugs.

Keywords:- *Drug-target interaction, CNN, Data balancing.*

I. INTRODUCTION

The DTI determination is a vast area of research which perform a crucial position in instigating advanced drugs for prior targets and for identifying advanced protein target for studied drugs. Number of potential untraced interactions were large and many of the drugs identified were refused due to its side effects or high toxicity. Conventional methods of laboratory experimentation were expensive and laborious. Computational approaches were developed to detect new DTIs. The public databases such as KEGG, DrugBank, etc,. which were developed based on verified interaction information, reserve and render knowledge on the grounds of laboratory experiments for constructing a computational application for determining latest DTIs and are applied as the gold standard dataset. To tackle limitations of previous models, a Convolutional Neural Network related model, deep ACTION is proposed to intoduce budding DTIs with the help of chemical structure and sequence information of drugs and proteins respectively. The pairs of drug-target and their corresponding chemical structures and protein sequences were downloaded from the DrugBank database and KEGG database.

Firstly, the chemical structure of drug is converted to a topo-logical, constitutional, and geometrical form. Various protein descriptors were utilized to describe sequence information of a target sequence. Then the valid data is created by combin-ing the extracted drug-protein features. Here the interacting features are indicated as positive pairs and non-interacting features as negative pairs. Then manage the imbalanced drug-target dataset using a data balancing technique and finally the training model for prediction of the interacting and non-interacting pairs were constructed by utilizing CNN classifier. Subsequently, when compared to other classifiers and methods, the suggested model exhibits the highest performance.

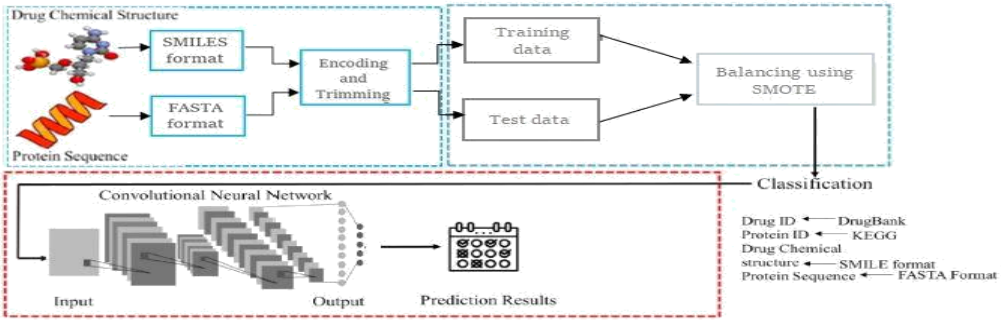

Fig. 1: Deep Action model

ISSN No:-2456-2165

II. DRUG TARGET INTERACTION (DTI)

The binding of a drug to a target position which concludes in an alteration in its behaviour or functionality is referred to as Drug target interactions. A drug or medicine primarily points to any chemical compound that could bring a physiological alterations in the human body when it is ingested, injected or sucked up. Target aka biological target can be any part of the living being to which the drugs get attached so as to introduce the physiological change. Entities like proteins or nucleic acids which are administered for any change are considered as targets. Nuclear receptors, ion channels, G-protein coupled receptors and enzymes are the most common biological targets.

DTI prediction takes part in an important position in the procedure of drug discovery. Drug discovery is focused in the identification of unidentified compounds of the drugs for different biological targets. Drug's chemical compound get attached to the molecules of target through bonds that are temporary. The chemical compounds of the drugs that are attached will then reciprocate with the biological target. This reaction may result in changes which can be positive or negative and accordingly draw out the target. In order to treat diseases, the drugs ward off certain catalyzed reactions happening in the human body by inhibiting the functioning of the target. It is attained by inhibiting the contact of drugs with substrates, which is a kind of enzyme.

There are two ways for the occurrence of the DTIs: One of the method is that, to impede the reaction the drugs which are known as competitive inhibitors will affix themselves to the active site of the target. Another method is that to avert the substrate from recognizing target. It is done by allosteric inhibitors which is a kind of drug which can affix to the target's allosteric site. Thereby it alters target's shape and structure. Thus, reactions will not take place. The shutting off of the reactions of the target can direct to the handling of metabolic imbalances or can kill pathogens in order to cure diseases.

DTI predictions has numerous applications. It can ease the procedures of drug discovery, repositioning of drug and prediction of side effects of drugs. Discovery of drugs can be defined as the inquiry of new drugs which is capable of interaction with a specified target.

III. BACKGROUND

A. Datasets

The dataset required for the experiment was freely available from DrugBank which is a Canadian online database. It contains information related to the drugs and targets. In the dataset, overall 12,674 interactions between drugs and targets are available. It comprises of 5877 drugs and 3348 target proteins. Distinct interpretations of the DrugBank dataset were utilized as standard datasets in earlier researches.

B. Extraction of Drug-target features

In the dataset, drugs are represented in SMILES format and targets in FASTA format. Simplified Molecular Input Line Entry System (SMILES) allows to describe the chemical structure of the drug in a way which can be utilized by the system. The chemical structure of the drug which is a 3-dimensional representation will be difficult to be manipulated by the system. So in order to ease the procedure of determi-nation of drugs the 3D representation is converted to a 1D representation. This resulting format is the SMILES format. FASTA is an acronym for FAST-ALL. It is an alignment tool. FASTA format is used to describe the nucleotide or amino acid sequences with single-letter codes. Specified drug and target IDs are used to gather the SMILES and FASTA format from DrugBank and KEGG databases respectively. These IDs are available from KEGG database. The extracted features are preprocessed.

As the generated features may exist in varied length, all of them are trimmed to fixed length vectors and it is then fed as an input to the chosen classifier. The features which are of no or least importance in the prediction procedure will be eliminated. In the end, an overall of 193 drugs and 1290 targets features are remained. It is then combined to form the drug-target pairs.

C. SMOTE For Balancing

In this proposed model, SMOTE technique which is an over-sampling method is included to get hold of the imbalanced datasets and to create the training dataset of drug-target pairs. For the sake of enhancing the prediction capability of the classifier by balancing the minority class, SMOTE syntheti-cally give rise to positive samples of the experimental datasets. SMOTE procedure is as follows:

- Initially select a sample randomly and from the minority class datasets detect the K-nearest neighbors for every samples in minority class
- The distance or dimension between the feature vector and its nearest neighbors are calculated
- Then the difference is multiplied using a random number from [0 to 1].
- This number is then added to the feature vector (sample)
- The process is continued until the minority and the majority classes have same number of samples and then stop the process

D. Convolutional neural network

A Convolutional neural network (CNN or ConvNet) is a kind of artificial neural network(ANN) and is employed in pattern recognition and image processing. It is particularly designed to exercise on pixel datas. They are recurrently utilized for various applications, such as analysis of voice, systems for recommendations, recognition and classification of videos and images and processing of natural language. ConvNet can reduce the images into a form that is convenient to process, without eliminating features that are crucial for making a good prediction. Generally, CNNs consist of an input layer, output layer and multiple hidden layers (MHLs). Multiple hidden layers in CNNs may include many convolutional layers (CLs). To alleviate the problem of model over fitting that may add noise to the HLs randomly, a dropout layer is assigned. Drop out layer is a regularization

ISSN No:-2456-2165

strategy. The nodes which are indicated as 'dropped out' will not either join in back propagation or assist in the forward pass. A common activation function called RELU layer is included and is accordingly observed by extra convolution layers.

In CNNs, main structure blocks are convolutional layers. For training more significant features deeper CLs are applied. This is achieved by including sliding kernels on the upper part of the earlier layers. Usually after each CL the pooling operation is applied. Thus, PLs can reduce the number of features provided and by local nonlinear functions can offer translation invariance .

IV. EXPERIMENTAL RESULTS

The suggested approach is entirely implemented with Python (version 3.6) as the programming language. Also incorporates Pytorch and scikit-learn library. Spyder, a free and open source environment which is written in Python is the em-ployed IDE. For the implementation on neural networks, Keras and tensorflow are included. Various tests were performed to examine the accuracy of various classifiers and balancing and selection approaches used. At the end, the classifier will put forward a list of the unused potential interactions.

A. Performance analysis

By setting parameters train the model on the basis of train-ing data. Some of the metrics for analyzing the performance include Accuracy, Sensitivity, Specificity and MCC . The AUC metric is used to evaluate the performance of deepACTION method. Then plot the ROC curve with TPR (sensitivity) in X-axis and FPR (1-specificity) in Y axiz by adjusting with different thresholds.

B. Performance analogy with other classifiers

A comparison with various classifiers on the Drug Bank dataset was performed to analyse the effectiveness and robust-ness of the deep ACTION model. Different classifier models obtain different accuracy values on same set of data.

The CNN classifier produced the highest AUC of 0.9133 for the dataset. GBN classifier achieved the second-highest result, with the performance AUC value of 0.8930. The classifiers Random Forest and KNN have performance AUC value of 0.885 and 0.882 respectively.

The CNN classifier have comparatively higher AUC values than the value obtained by the classifiers GBN, Random Forest and KNN. By comparing the AUC values, it is clear that the accuracy of CNN is 2.03%, 2.83%, and 3.13% higher than the other classifiers used respectively.

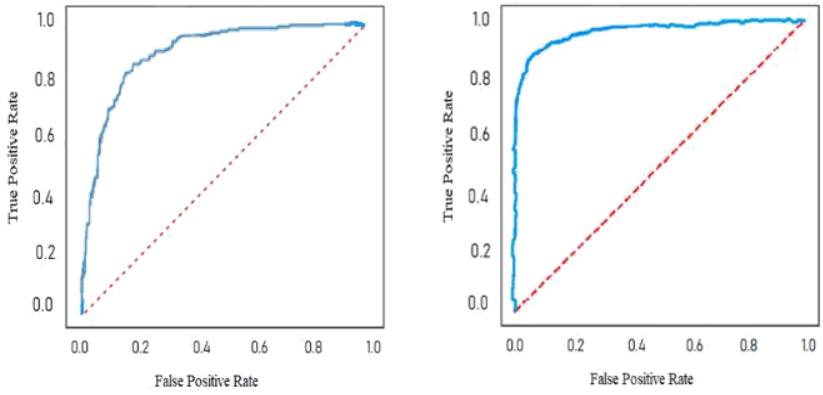

Fig. 2: Performace Evaluation: Accuracy

```
CNN

CNN Acuracy is : 91.13333225250244 %
              precision    recall  f1-score   support

         0.0       0.50      1.00      0.67      4500
         1.0       0.00      0.00      0.00      4500

    accuracy                           0.50      9000
   macro avg       0.25      0.50      0.33      9000
weighted avg       0.25      0.50      0.33      9000

Confusion Matrix:
[[4500    0]
 [4500    0]]
```

Fig. 3: ROC Curve Before and After Balancing Using SMOTE

ISSN No:-2456-2165

V. FUTURE RESEARCH

Presently the project uses a SMILES format to represent drugs and FASTA format to represent proteins. As a future work, it is possible to utilize a ligand-based protein representation method that uses SMILES sequences of the interacting ligands to describe proteins. So that both the input dataset can be represented in the same format.

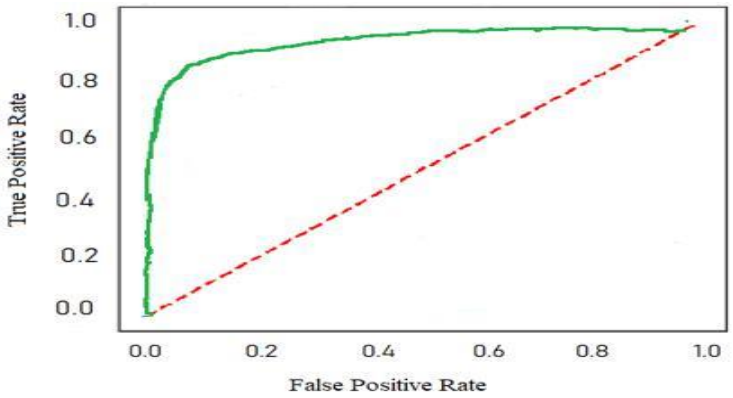

Fig. 4: ROC Curve of GBN

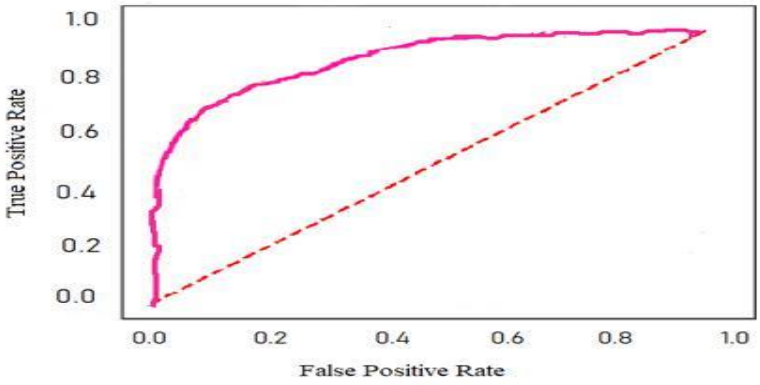

Fig. 5: ROC Curve of KNN

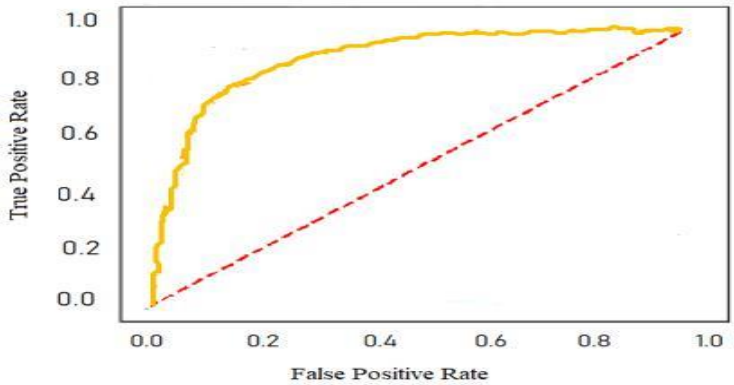

Fig. 6: ROC Curve of RandomForest

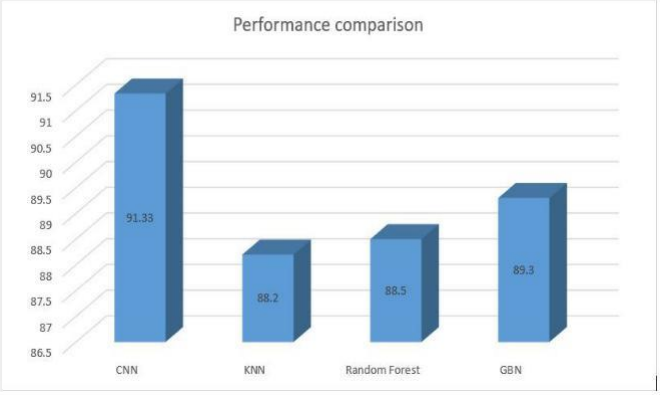

Fig. 7: Performance comparison of different classifiers

Also, a drug-target dataset (networks) which is heteroge-neous with auPR matrices for experiments can be considered and a stand-alone web application can be developed by pro-viding simple mechanisms and a user-friendly interface for the model. In future, research can also be extended to predict related directions such as drug-drug, drug side effects, and protein diseases.

By integrating multiple heterogeneous data sources of drugs and targets, a more efficient prediction model can be created.

VI. CONCLUSION

The project developed a model based on deep learning called deepACTION for the prediction of interactions between drugs and targets . Interactions between drugs and targets are determined using CNN algorithm. The drug-target structure is represented in numeical form by utilizing various feature extraction techniques. Related works used different methods to manage the imbalanced datasets. DeepAction model include SMOTE to manipulate the majority (negative) and minority (positive) instances in the dataset. SMOTE balance class distribution by randomly increasing minority class samples by replicating them.

Different classifiers such as CNN, GBN, Random Forest and KNN were applied. CNN returned higher accuracy of 91.33% compared to other classifiers. GBN classifier returns the second-highest accuracy of 89.30%. Random Forest and KNN produced accuracy of 88.5% and 88.2% respectively.

From the experimental results it is evident that the deep-Action method attains highest prediction performance and is capable to predict novel drug-target pairs from the DrugBank dataset. The improved performance of proposed model may motivate scientists to utilize this method in prediction of new DTIs.

REFERENCES

[1.] S.M. Hasan Mahmuda, Wenyu Chena, Hosney Jahanb, Bo Daia, Salah Ud Dina, Anthony Mackitz Dzisoo," DeepACTION: A deep learning-based method for predicting novel drug-target interactions", October 6,2020..

[2.] ShanShan Hu, DeNan Xia, Benyue Su, Peng Chen , Bing Wang , and Jinyan Li,"A Convolutional Neural Network System to Discriminate Drug-Target Interactions", August 2021. .

[3.] Ping Xuan , Bingxu Chen, Tiangang Zhang , and Yan Yang, "Prediction of Drug–Target Interactions Based on Network Representation Learning and Ensemble Learning", December 2021.

[4.] ABDELRAHMAN I. SAAD , YASSER M. K. OMAR, AND FAHIMA

[5.] MAGHR, "Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques", October 22, 2019.

[6.] SKonstantinos Pliakos , Celine Vens , and Grigorios Tsoumakas, "Pre-dicting Drug-Target Interactions With Multi-Label Classification and Label Partitioning", August 2021.

[7.] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan and S. Ahmed, "iDTi-CSsmoteB: Identification of Drug–Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost With Over-Sampling Technique SMOTE," in IEEE Access, vol. 7, pp. 48699-48714, 2019, doi: 10.1109/ACCESS.2019.2910277.

[8.] M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, P. Bork, Drug target identification using side-effect similarity, Science 321 (2008) 263–266, https://doi.org/10.1126/ science.1158140, 80-.

[9.] H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou,

[10.] Wang, Y. Wang, A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data, PLoS One 7 (2012), https://doi.org/10.1371/

[11.] S. Hu, C. Zhang, P. Chen, P. Gu, J. Zhang, B. Wang, Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks, BMC Bioinf. 20 (2019) 1–12, https://doi.org/10.1186/s12859- 019-3263-x.

[12.] Ezzat, M. Wu, X.L. Li, C.K. Kwoh, Drug-target interaction prediction via class imbalance-aware ensemble learning, BMC Bioinf. 17 (2016), https://doi.org/ 10.1186/s12859-016-1377-y.
[13.] S. Hu, D. Xia, B. Su, P. Chen, B. Wang, J. Li, A convolutional neural network system to discriminate drug-target interactions, IEEE ACM Trans. Comput. Biol. Bioinf (2019), https://doi.org/10.1109/TCBB.2019.2940187.
[14.] T. Pahikkala, A. Airola, S. Pietila, Toward more realistic drug-target interaction predictions, Briefings Bioinf. (2014) 1–13, https://doi.org/10.1093/bib/bbu010.