Importance of Drug Features in Drug-Drug Interaction : A Comparative Study

Chetana N Patil
Department of CSE
PES University
Bengaluru, India
chetananpatil2002@gmail.com

Naveen B N
Department of CSE
PES University
Bengaluru, India
naveenbiligeri@gmail.com

Mekala Sanjana
Department of CSE
PES University
Bengaluru, India
sanjana.mekala135@gmail.com

Kumari Shivangi
Department of CSE
PES University
Bengaluru, India
shivangik325@gmail.com

Prajwala T R
Associate Professor Department of CSE
PES University
Bengaluru, India
prajwalatr@pes.edu

***Abstract*— A Drug-Drug Interaction is an interplay of drugs where one or more drugs interfere with the activity of other drugs. Synergy or Interaction between drugs may cause side effects that are unforeseen.They may also intervene and oppose each other's action and nullify the medication.Many of such interactions are often negligible but some might be harmful if not discovered at the earliest.Features or characteristics of drugs do play a vital role in interaction.It therefore becomes an essential study to untie the patterns of interaction.**

Our study aims to predict important features that are involved in the interaction of two drugs.Firstly we train multiple learning models on the dataset. These models are RandomForestClassifier,XGBoost,SVM,Autoencoders+XGBoost,CNN and Capsule Networks. Then compare the results of trained models and select the most appropriate performance yielding model. We then ask the model to predict the interaction between two drugs and also plot the feature importance distribution of two interacting drugs. In the study it was found that RandomForestClassifier did slightly better than XGBoost and outshined against all other models.Hence RandomForestClassifier was used as The BackBone model to predict feature importances amongst two interacting drugs.

***Keywords*—Drug-Drug Interaction, Random Forest Classifier,XGBoost,CNN,SVM,Capsule Networks, Autoencoders**

I. INTRODUCTION

In recent years AI driven drug discovery is advancing as scientific innovations in this field can reduce the number of years taken to develop and design the drugs. People all across the world, especially the elderly people ingest multiple drugs at a time for underlying medical causes due to the pervasiveness of polypharmacy[2].And these drugs when ingested simultaneously together might lead to an interaction either with another drugs or with a protein or any target molecules and might lead to Adverse Side Reactions(ADR) ADR are one of the main reasons for financial loss in the pharmaceutical company as ADR's might cause mortality and morbidity and ultimately leading to the withdrawal of the drugs sold across the markets[4].Therefore has paved a way for the discovery of these interactions .

As we all know that there could be multiple DrugDrug Interaction can take place.In our study we restricted ourselves to find the interaction taking place between two drugs .Each drug is characterized by its features namely permeability,solubility ,structural pattern etc.These all features tend to examine four major parameters of ingested drugs:absorption,distribution,metabolism and Excretion in short ADME's . It therefore becomes an essence to identify to study the interaction patterns that take place between two drugs by using these properties.

If we take any two drugs that are interacting , not all attributes of source and target drug molecule influence its importance in the interaction.Hence,it is vital for us to study these interaction patterns to predict the feature importances in the Drug-Drug Interaction.

We aim to build multiple models and then select the best model based on performance and then use that model to predict the importance of Drug features.The study not only help drug designers to develop new drugs but also help them rethink on the solution for repurposing some combination of drugs if they are found to be synergistic.This would then ultimately lead efficient and proficient Drug discovery.

II. LITERATURE REVIEW

3WDDI is a novel method for predicting Drug Drug interactions using a 3 way decision process as well as the application of the knowledge graph. Here CNN was used as a delay decision model and used as a decision function.They generated drug substructure similarity features and drug embeddings from the Drug Knowledge Graph (DRKG) using ComplEx for the Three-Way Drug-Drug Interaction (3WDDI)

prediction. Subsequently, the drug substructure similarity features were employed to delineate the boundary region and derive classification results for the remaining areas. The knowledge graph embeddings of a drug pair were introduced as novel supplementary features by 3WDDI for improved delay decision-making. The proposed method was implemented, and a comparative experiment was conducted on a widely adopted dataset. The experimental outcomes demonstrate the superior performance of 3WDDI compared to baseline Drug-Drug Interaction (DDI) prediction models.[1]

CNN-DDI, a refined architecture of CNN, was used to predict Drug interactions based on similarity.In this method features were extracted from various descriptors of drugs namely,drug categories,enzymes and pathways.Then they converted it into feature vectors and then spatial information amongst the vectors were obtained using Jaccard Similarity.Then based on the similarity metrics obtained ,they developed a new Convolutional Neural Network Model.But the accuracy of the model was 88.[3]

DDIPred was a model developed to find the importance of chemical structure of drugs in the interaction.The study involved use of SMILES(Simplified Molecular Input Line Entry System).Susceptibility of syntactic and semantic error is plausible in SMILES and hence was converted to SELFIES(SELFreferencIng Embedded Strings) to study the interaction importance.Graph Convolutional Network was used as a backbone to study the interaction.The method had good AUROC of 0.99.However the model studied only one feature that is Drug's chemical structure.[4]

Multiple machine learning methods have been used to predict drug-target interactions. To evaluate the proposed models, the calculated AUROC and AUPRC values and obtained results indicate that nonlinear SVM and logistic regression performed better than other models with AUROC and AUPRC values of 0.8317 and 0.8260, respectively.The cons of the model are: It may lead to overfitting. Logistic regression is not suitable for modeling complex relationships between variables, such as interactions or nonlinear effects. Logistic regression is limited to linear relationships. [6].

DANN DDI , a model developed using Deep Attention Neural Networks.Here in the model they attentively added the drug features to predict unforeseen drug drug interactions.Accuracy score of 88.74 was obtained in the overall model.Proper balanced data and proper noise removal in the data would lead to improved accuracy scores.[12]

When Deep Neural Network was used as a model to predict the interaction of drugs, accuracy of 93.2 on test dataset and 94.9 on validation dataset were obtained.This method involved small dataset of 5,134 drug molecules with their drug descriptors namely constitutional properties,topological properties ,geometric properties.The model used four hidden layers and activation unit as RELU.Further more on the study they added the testing to predict interaction amongst existingIBD drugs and found an increase in antipsychotic effects. of drugs.[14]

Below are some of the conclusion that were made from the entire literature Review :

- Majorly the ML and DL model used were SVM , Random Forest And Decision Tree, CNN, Autoencoders.
- They have compared the performances of these algorithms using F-1 score , accuracy,precision and AUC.
- We used ML algorithms to predict the drug interaction and the most important feature responsible for the interaction.

III. DATASET

Number of dataset that facilitated our study was three.

1. Drug-Drug Interaction Dataset
2. Drug features Dataset
3. Drug Names Dataset

A. Drug-Drug Interaction Dataset This dataset represents the presence or absence of interaction between two drugs.The columns present here include 'source', 'target' and 'interaction'. The 'interaction' column consists of two values 0 and 1.Value ' 0' indicates no interaction exists between the two drugs and Value '1' indicates the presence of interaction between 'source' and the 'target' columns as shown in Fig 1.

	source	target	interaction
0	DB00862	DB00966	1
1	DB01235	DB01275	1
2	DB01609	DB06212	1
3	DB01232	DB09291	1
4	DB00104	DB00908	1
...	...	...	...
22967	DB00632	DB15299	0
22968	DB00633	DB15300	0
22969	DB00634	DB15302	0
22970	DB00635	DB15307	0
22971	DB00640	DB15308	0

Fig. 1. Drug-Drug Interaction Dataset

B. Drug features Dataset The refined dataset comprises 22972 rows.This dataset represents features of individual drugs.The dataset spans over 8288 rows and 7 columns. The columns include 'drugbank id' and 'cas' drug identification Id followed by these columns are the features representing permeability and solubility attributes of drugs namely 'log ALOGPS','logP

hemAxom, 'solubility', 'LOGPS', 'pKa(strongest acidic)' and 'pKa(strongest basic)'features as seen in Fig 2 .

	drugbank_id	cas	logP ALOGPS	logP ChemAxon	solubility ALOGPS	pKa (strongest acidic)	pKa (strongest basic)
0	DB00006	128270-60-0	-0.76	-14.00	4.64e-02 g/l	2.79	11.88
1	DB00007	53714-56-0	1.04	-2.40	3.38e-02 g/l	9.49	11.92
2	DB00014	65807-02-5	0.30	-5.20	2.83e-02 g/l	9.27	10.82
3	DB00035	16679-58-6	-1.00	-6.10	1.10e-01 g/l	9.50	11.77
4	DB00050	120287-85-6	1.33	-1.70	6.94e-03 g/l	9.50	11.79

Fig. 2. Drug Features Dataset

C. Drug Names Dataset This data comprises three columns namely, 'drugbank id', 'name' and 'SMILES' columns.'name' column represents the name of the drug corresponding to 'drugbank id' as seen in Fig 3.

	drugbank_id	name	smiles
0	DB00006	Bivalirudin	CC[C@H](C)[C@H](NC(=O)[C@H](CCC(O)=O)NC(=O)[C@...
1	DB00007	Leuprolide	CCNC(=O)[C@@H]1CCCN1C(=O)[C@H](CCCNC(N)=N)NC(=...
2	DB00014	Goserelin	CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@H...
3	DB00035	Desmopressin	NC(=O)CC[C@@H]1NC(=O)[C@H](CC2=CC=CC=C2)NC(=O)...
4	DB00050	Cetrorelix	CC(C)C[C@H](NC(=O)[C@@H](CCCNC(N)=O)NC(=O)[C@H...

Fig. 3. Drug Names Dataset

IV. METHODOLOGY

Proposed overflow involves four major steps,

1.Data Preprocessing

2.Application of Models

3.Selection of Model

4.Feature Importance Prediction

A. Dataset Preprocessing Individual preprocessing of the dataset is carried out.

Step 1 : Removal of Unwanted Features:

We proceed with use of 'drugbankid' as our reference id for the identification of drug and 'cas' drug identifier column is unnecessary and hence is removed from Drug Features Dataset.'SMILES' column present in Drug Names dataset provides no use to our estimation and is eliminated.

Step 2 : Joining of datasets:

First part of this step includes inner join on two datasets i.e, Drug Features Dataset and Drug Names Dataset on the column 'drugbank id'. New dataset formed is named as the Drug Feat Names dataset. In the next part we join the 'source' column from the Drug-Drug Interaction dataset with the 'drugbank id' column of Drug Feat Names dataset. Results yielding from this join are saved as Source Dataset. We rename columns from 'drugbankid', 'logALOGPS', 'logPChemAxom', 'solubility LOGPS', 'pKa(strongest acidic)' and 'pKa(strongest basic)' to 'source drugbank id', 'source logP', 'source logP CA', 'source solubility', 'source acidic pKa' and 'source basic pKa'. Similarly we tend to join the 'target' column from the Drug-Drug Interaction dataset with the 'drugbank id' column of Drug Feat Names dataset.Results yielding from this join are saved as Target Dataset. We rename columns from 'drugbankid', 'logALOGPS', 'logPChemAxom', 'solubility LOGPS', 'pKa(strongest acidic)' and 'pKa(strongest basic)' to 'target drugbank id', 'target logP', 'target logPCA', 'target solubility', 'target acidic pKa' and 'target basic pKa'. As a part of the last step we merge Source Dataset and Target Dataset based on Drug Drug Interaction dataset and save as Merged Dataset.This dataset comprises of columns 'source drugbank id', 'source logP', 'source logP CA', 'source solubility', 'source acidic pKa', 'source basic pKa', 'target drugbank id', 'target logP', 'target logP CA', 'target solubility', 'target acidic pKa' and 'target basic pKa'.

Step 3 : Removal of Null values

Any row entries having NA values are removed in this step.After this step our dataset is ready to be utilized by the models.

B. Models

1) RandomForestClassifier: RandomForestClassification technique creates multiple decision tree estimators by the utilization of different dataset subsets.We split the dataset for training and testing by 80:20 ratio.And then train the model using 100 tree estimators to output the predictions.

2) XGBoost: XGBoost is the most powerful form of gradient boosting techniques.The model comprises decision tree estimators analyzed sequentially rather than parallel as in case of RandomForestClassification. Here again we make use of 100 decision tree estimators for classification and 80 data for training.

3) XGBoost + Autoencoders: This model takes the predictions from XGBoost and combines with the encoded features obtained from the autoencoders.A simple classifier layer is trained above the encoded features . Here again data is split in the ratio 80:20.

4) SVM : SVM draws a hyperplane separating classes .Here the hyperplane studies the drug features and solely separates into two parts i.e,the drug classes that interact and the ones which do not.

5) CNN: Input layer receives features and passes down to two layers of convolution and a flatten layer to extract essential features. Activation unit used in the Convolution layer is RELU.A dense layer added on top of it for classification purposes. Activation unit used in the dense layer is

Sigmoid.Then we compile the model with binary cross entropy loss, Optimiser as Adam and activation unit as RELU and it is made to run for 20 epochs to obtain desirable results.

6) Capsule Networks: Capsule Networks are one of the widely used Deep learning technique.This model comprises three layers. Input layer reads the features and is passed to the first layer of Capsule Network i.e. Primary Capsule layer which extracts essential features in the form of vectors. Followed by that extracted features are sent to Digit Capsule Layer where Squash function is used to extract the spatial relationship amongst the vector data.The relationship learnt in this layer is utilized by Class Capsule layer to predict the results.

Function squash(vectors):
squared norm = sum(square(vectors)) scale = squared norm / (1 + squared norm) / sqrt(squared norm + epsilon) squashed vectors = scale * vectors

Fig. 4. Flow of activities

C. Selection of model:
Metrics of performance are calculated on each of the models we discussed above for selection of good result yielding model.Metrics studied include Accuracy score, Precision Score , Recall Score , F1 score , Mean Squared Error (MSE) and ROC- AUC values.
The performance estimation of any classification model can be judged by Receiver Operating Curve(ROC). It gives Area Under the Curve (AUC) value which helps in the estimation of performance. From Fig1 to Fig5 it is evident that RandomForestClassifier has yielded better results with AUC 0.96 which is slightly ahead of XGBoost's AUC- 0.94.

Algorithm	Accuracy	Precision	Recall	F1 Score	MSE	ROC-AUC
Random Forest	95.60	95.63	99.83	0.97	0.03	0.96
XG-Boost	95.71	95.88	99.67	0.97	0.04	0.94
SVM	93.32	93.31	99.97	0.96	0.06	0.67
Capsule Networks	92.90	92.97	99.92	0.96	0.10	0.67
CNN	93.61	94.03	99.44	0.96	0.05	0.77
XG-Boost + Auto-encoder	93.30	94.42	98.61	0.96	0.06	0.61

Fig. 5. Metrics of Performance

Now let us look at other plots for comparison i.e,.,Precision Recall Curve which attributes to estimate the trade offs that exist between True Positive Rate and actual predicted positive value.

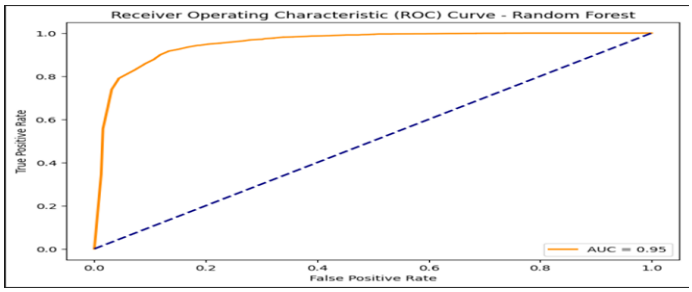

Fig. 6. ROC-RandomForestClassifier

The Fig6 represents the ROC Curve of Random forest. AUC of 0.95 indicates a very good model performance. i.e; it has a high ability to distinguish between positive and negative instances.

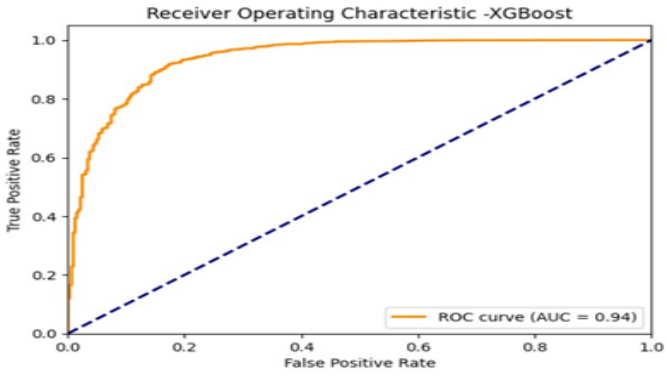

Fig. 7. ROC - XGBoost

The Fig7 represents the ROC Curve of XGBoost.With an AUC of 0.94 this model is also a very good model like Random Forest Classifier. i.e, it has a high ability to distinguish between positive and negative instances.

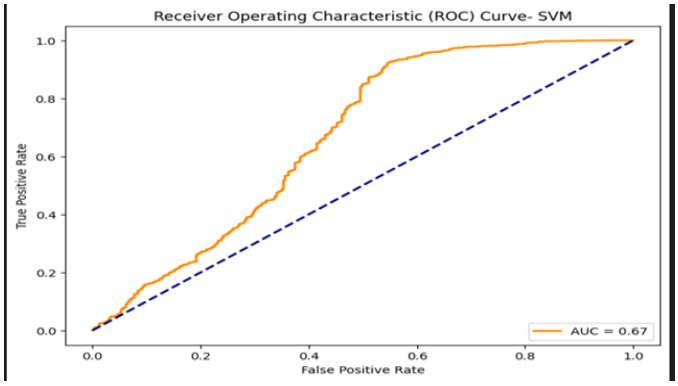

Fig. 8. ROC - SVM

The Fig8 represents the ROC Curve of SVM. AUC of 0.67 indicates moderate performance. i.e, there is a good balance between true positive rate and false positive rate.

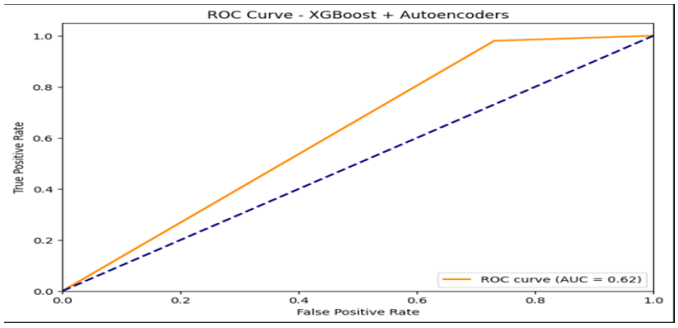

Fig. 9. ROC - XGBoost+Autoencoder

The Fig9 represents the ROC Curve of XGBoost+Autoencoders. AUC of 0.62 indicates lower performance compared to all the other models.

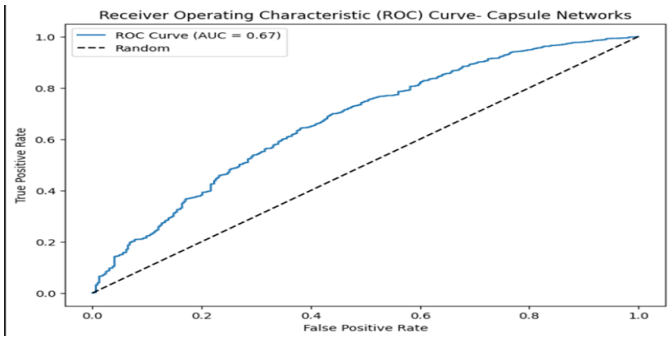

Fig. 10. ROC - Capsule Network

The Fig10 represents the ROC Curve of Capsule Network. AUC of 0.67 is similar to SVM suggesting comparable performance.

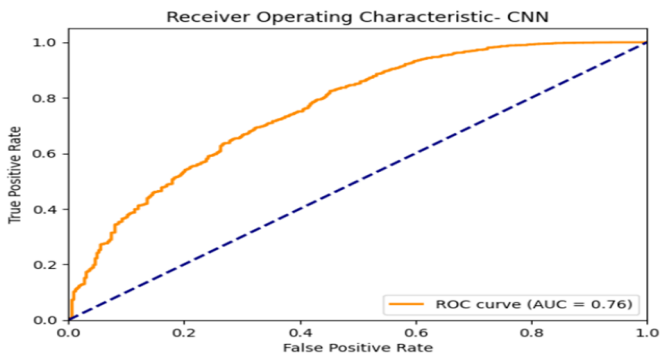

Fig. 11. ROC - CNN

The Fig11 represents the ROC Curve of CNN.AUC of 0.76 suggests that the model performed reasonably well i.e; there is a good balance between true positive rate and false positive rate.

By looking at the plots Fig 6 to Fig 11,again it is evident that RandomForest Classifier outshines other models. If we debate on the Mean Squared Error(MSE) values it it evident that RandomForestClassifier has the lowest value of 0.03 followed by XGBoost with 0.04 ,CNN with 0.05 , SVM and XGBoost+Autoencoders with 0.06 and lastly Capsule Network with 1.0. Hence we conclude that RandomForestClassifier is the model to be chosen for the Feature Importance analysis.

D. Feature Importance Estimation:

We make use of the trained RandomForest model above and add upon an interface for predicting and testing the results.Tkinter library of python was used to build GUI. The GUI checks firstly whether an interaction exists between the drugs and if at all the interaction is found out then it plots the feature importance of the two interacting drugs. The GUI also outputs the highest feature importance.

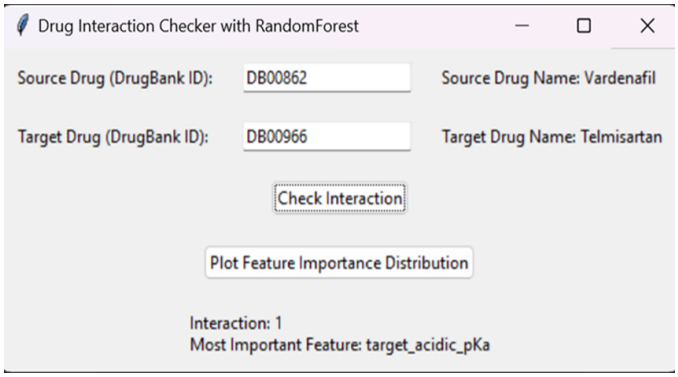

Fig. 12. GUI for interaction between drugs

Fig 12 depicts the interface developed using tkinter.It takes two inputs Source DrugBank ID and Target DrugBank ID.We

input the drugbank id's in the field. RandomForest Model trained on the dataset identifies the presence of interaction and outputs 1 if interaction exists between the drugs when clicked on 'Check Interaction' button.It also outputs their respective names and even the highest prevalent feature Importance.

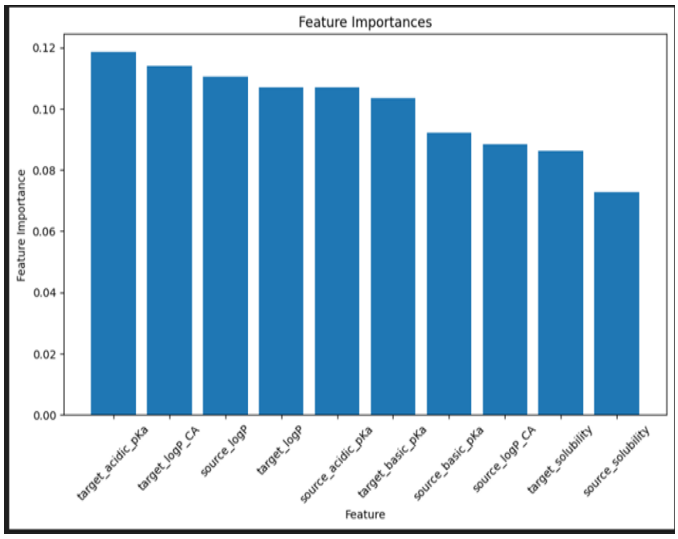

Fig. 13. Feature Importance Distribution

Next when clicked on 'Plot Feature Importance Distribution' the above graph Fig 13 is depicted on the screen.

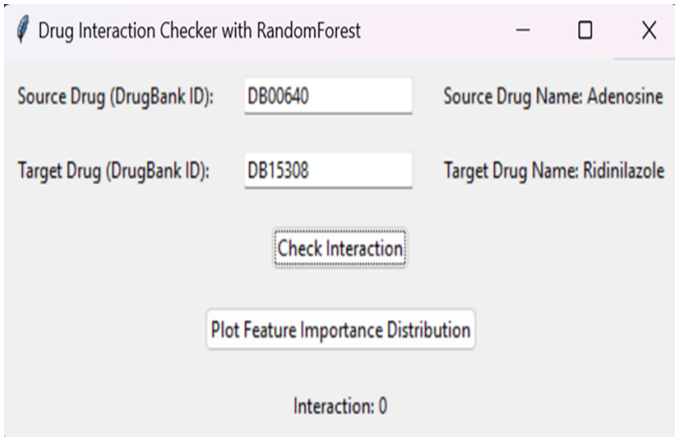

Fig. 14. GUI for no interaction between drugs

Fig 14 depicts two drugs posing no interaction when clicked on 'Check Interaction' button.If we Click on the button 'Plot feature Importance Distribution' button error message is displayed on the terminal 'Feature Importance graph exists only if there is an interaction.

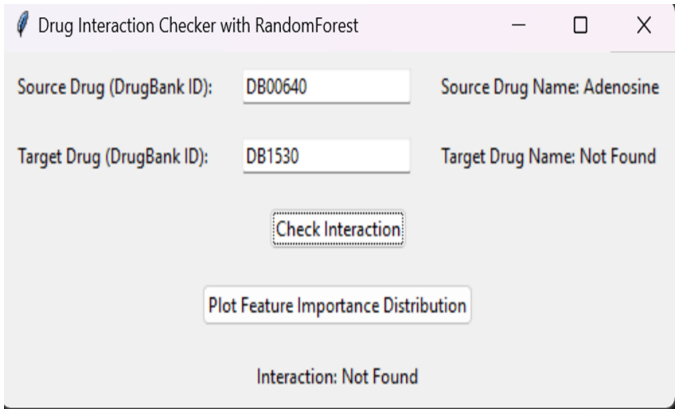

Fig. 15. GUI for interaction not found between drug

Fig 15 shows incorrect Target DrugBank Id which when clicked on 'Check Interaction' button displays that Interaction is not found.When clicked on 'Plot Feature Importance Distribution' error message will be displayed on the terminal showing 'No data found'.

V. RESULTS

The project yielded promising results. Accuracy metrics for the models showed strong performance, with Random Forest achieving an accuracy of 95.60 percentage, Support Vector Machine(SVM) achieving an accuracy of 93.3 percentage, XGBoost achieving an accuracy of 95.4 percentage.The system uses Random Forest algorithms to analyze interaction between drugs based on drug features and gives output of most responsible feature for interaction.

VI. CONCLUSION AND FUTURE WORK

The goal of the project was to train multiple models on drug-drug interactions. Evaluate and select the best model by comparing models performance metrics to predict the importance of features of two interacting drugs. We all know that not only these five major features of drugs are contributing to Drug-Drug Interaction there are numerous other attributes of drugs that influence the interaction.Our study was limited to two drugs and five major features but in real time multiple Drug-Drug Interaction could take place.In our future projects we aim to increase the number of features involved in the interaction as well as introduce the scope of learning multiple drug-drug interactions.

REFERENCES

[1] Xinkun Hao, Qingfeng Chen, Haiming Pan,Jie Qiu, Yuxiao Zhang,Qian Yu,Zongzhao Han,Xiaojing Du,"Enhancing drug–drug interaction prediction by three-way decision and knowledge graph embedding',Springler Nature,2022

[2] E. Kim and H. Nam, "Decide-ddi: interpretable prediction of drug drug interactions using drug-induced gene expressions," Journal of cheminformatics, vol. 14, no. 1, pp. 1–12, 2022

[3] ChengCheng Zang, Yao Lu ,Tianyi Zang ,"CNN-DDI: a learning-based method for predicting drug–drug interactions using convolution neural networks",BMC Bioinformatics,23,Article number :88(2022)

[4] Shaghayegh Sadeghi , Alioune Ngom, "DDI PRED:Graph Convolutional Network based Drug-Drug Interactions Prediction using Drug Chemical Structure Embedding" in IEEE, 2022.

[5] Nilay Fatma , Yıldız Alper Ozcan, "Graph Convolutional Autoencoder ¨ and Generative Adversarial Network-Based Method for Predicting DrugTarget Interactions " published in 2022.

[6] Xu Gong, Maotao Liu, Haichao Sun,Min Li, Qun Liu, "HS-DTI: DrugTarget Interaction Prediction based on Hierarchical Networks and MultiOrder Sequence Effect" published in 2022.

[7] Tianjung , Wang , Xin , Liu, "A Graph Convolution-Transformer Neural Network for Drug-Target Interaction Prediction", ICBBT 2022, May 27–29, 2022.

[8] Chang Sun, Ping Xuan , Tiangang Zhang and Yilin Ye, "Using Supervised Machine Learning Algorithms for Drug-Target Interaction Prediction" in IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 19, No. 1, January/February 2022.

[9] Mohammad A. Rezaei , Yanjun Li , Dapeng Wu, Xiaolin Li and Chenglong Li,"Deep Learning in Drug Design : Protein-Ligand Binding Affinity Prediction" in IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume: 19, Issue: 1, 01 Jan.-Feb. 2022.

[10] Nelson R. C. Monteiro , Bernardete Ribeiro , and Joel P. Arrais,"Drug Target Interaction Prediction : end-to-end Deep Learning Approach" in IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume: 18, Issue: 6 , 01 Nov.-Dec. 2021.

[11] Liu S., An J., Zhao J., Zhao S., Lv, H. Wang S, "Drug-Target Interaction Prediction based on Multisource Information Weighted Fusion" from Drug-Target Interaction Prediction Based on Multisource Information Weighted Fusion.Contrast media molecular imaging, 6044256, 2021.

[12] Shichao Liu, Yang Zhang, Yuxin Cui, Yang Qiu, Yifan Deng, Wen Zhang, Zhongfei Zhang, "Enhancing Drug-Drug Interaction Prediction Using Deep Attention Neural Networks ", IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS,2021.

[13] Mongia A, Majumdar A,"Drug-Target Interaction Prediction using Multi Graph Regularized Nuclear Norm Minimization", from PLoS ONE 15(1): e0226484, 2020.

[14] Xinyu Hao , Jiaying You , PingZhao Hu, "Predicting Drug-Drug Interactions using Deep Neural Networks", ICMLC '19, February 22–24, 2019..