The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 19(2): S2: 318-322, 2024 www.thebioscan.com

DEEPSIDE A DEEP LEARNING FRAMEWORK FOR DRUG SIDE EFFECT PREDICTION

[1]JAKKAMSETTY NAGADURGALAKSHMI, [2]NETHALATULASI RAJU,

[3] DR. P. SRINIVASULU

[1]PG Student, Department of CSE, Swarnandhra College of Engineering & Technology, Narsapur, AP, India.

[2]Associate professor, Department of CSE, Swarnandhra College of Engineering & Technology, Narsapur, AP, India.

[3]Professor & HOD, Department of CSE, Swarnandhra College of Engineering & Technology, Narsapur, AP, India.

[1]durga.jakkamsetti27@gmail.com , [2]raju.tulasi81@gmail.com ., [3]drspamidi@gmail.com

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S2.pp318-322

Received on:

25-07-2024

Accepted on:

07-11-2024

ABSTRACT

Drug failures due to unforeseen adverse effects at clinical trials pose health risks for the participants and lead to substantial financial losses. Side effect prediction algorithms have the potential to guide the drug design process. LINCS L1000 dataset provides a vast resource of cell line gene expression data perturbed by different drugs and creates a knowledge base for context specific features. The state-of-the-art approach that aims at using context specific information relies on only the highquality experiments in LINCS L1000 and discards a large portion of the experiments. In this study, our goal is to boost the prediction performance by utilizing this data to its full extent. We experiment with 5 deep learning architectures. We find that a multi-modal architecture produces the best predictive performance among multi-layer perceptron-based architectures when drug chemical structure (CS), and the full set of drug perturbed gene expression profiles (GEX) are used as modalities. Overall, we observe that the CS is more informative than the GEX. A Convolutional neural network-based model that uses only SMILES string representation of the drugs achieves the best results and provides 13:0% macro-AUC and 3:1% micro-AUC improvements over the state-of-the-art. We also show that the model is able to predict side effect-drug pairs that are reported in the literature but was missing in the ground truth side effect dataset.

INTRODUCTION

Computational methods hold great promise for mitigating the health and financial risks of drug development by predicting possible side effects before entering into the clinical trials. Several learning based methods have been proposed for predicting the side effects of drugs based on various features such as: chemical structures of drugs [25, 1, 23, 8, 19, 34, 17, 9, 2, 5], drug-protein interactions [35, 33, 8, 19, 34, 17, 37, 2, 15, 36], protein-protein interactions (PPI) [8, 9], activity in metabolic networks [38, 26], pathways, phenotype information and gene annotations [8]. In parallel to the above mentioned approaches, recently, deep learning models have been employed to predict side effects: (i) [31] uses biological, chemical and semantic information on drugs in addition to clinical notes and case reports and (ii) [4] uses various chemical fingerprints extracted using deep architectures to compare the side effect prediction performance. While these methods have proven useful for predicting adverse drug reactions (ADRs – used\ interchangeably with drug side effects), the features they use are solely based on external knowledge about the drugs (i.e., drug-protein interactions, etc.) and are not cell or condition (i.e., dosage) specific. To address this issue, Wang et al. (2016) utilize the data from the LINCS L1000 project [32]. This project profiles gene expression changes in numerous human cell lines after treating them with a large number of drugs and small-molecule compounds. By using the gene expression profiles of the treated cells, [32] provides the first comprehensive, unbiased, and cost-effective prediction of ADRs. The paper formulates the problem as a multi-label classification task. Their results suggest that the gene expression profiles provide context-dependent information for the side-effect prediction task. While the LINCS dataset contains a total of 473,647 experiments for 20,338 compounds, their method utilizes only the highest quality experiment for each drug to minimize noise. This means that most of the expression data are left unused, suggesting a potential room for improvement in the prediction performance. Moreover, their framework performs feature engineering by transforming gene expression features to enrichment vectors of biological terms. In this work, we investigate whether the incorporation of gene expression data along with the drug structure data can be leveraged better in a deep learning framework without the need for feature engineering. In this study, we propose a deep learning framework, Deep Side, for ADR prediction. Deep Side uses only (i) in vitro gene expression profiling experiments (GEX) and their experimental meta data (i.e., cell line and dosage - META), and (ii) the chemical structure of the compounds (CS). Our models train on the full LINCS L1000 dataset and use the SIDER dataset as the ground truth for drug - ADR pair labels [13]. We experiment with five architectures: (i) a multi-layer perceptron (MLP), (ii) MLP with residual connections (Res MLP), (iii) multi-modal neural net- works (MMNN. Concat and MMNN. Sum), (iv) multi-task neural network (MTNN), and finally, (v) SMILES convolutional neural network (SMILES Conv). We present an extensive evaluation of the above-mentioned architectures and investigate the contribution of different features. Our experiments show that CS is a robust predictor of side effects. The base MLP model, which uses CS features as input, produces _11% macro-AUC and _2% micro- AUC improvement over the state-of-the-art results provided in [32], which uses both GEX (high quality) and CS features. The multi-modal neural network model, which uses CS, GEX and META features and uses summation in the fusion layer (MMNN. Sum) achieves 0:79 macro-AUC and 0:877 micro-AUC which is the best result among MLP based approaches. We also find out

that when the chemical structure features are fully utilized in a complex model like ours, it overpowers the information that is obtained from the GEX dataset. The Convolutional neural network that only uses the SMILES string representation of the drug structures achieves the best result among all the proposed architectures with provides 13:0% macro-AUC and 3:1% micro-AUC improvement over the state-of-the-art algorithm. Finally, inspecting the confident false positives predictions reveal side effects that are not reported in the ground truth dataset, but are indeed reported in the literature. Deep Side is implemented and released at http://github.com/OnurUner/DeepSide.

2. LITERATURE SURVEY

1) Drug Side Effect Prediction with Deep Learning Molecular Embedding in a Graph-of-Graphs Domain

Abstract: Drug side effects (DSEs), or adverse drug reactions (ADRs), constitute an important health risk, given the approximately 197,000 annual DSE deaths in Europe alone. Therefore, during the drug development process, DSE detection is of utmost importance, and the occurrence of ADRs prevents many candidate molecules from going through clinical trials. Thus, early prediction of DSEs has the potential to massively reduce drug development times and costs. In this work, data are represented in a non-euclidean manner, in the form of a graph-of-graphs domain. In such a domain, structures of molecule are represented by molecular graphs, each of which becomes a node in the higher-level graph. In the latter, nodes stand for drugs and genes, and arcs represent their relationships. This relational nature represents an important novelty for the DSE prediction task, and it is directly used during the prediction. For this purpose, the MolecularGNN model is proposed. This new classifier is based on graph neural networks, a connectionist model capable of processing data in the form of graphs. The approach represents an improvement over a previous method, called DruGNN, as it is also capable of extracting information from the graph-based molecular structures, producing a task-based neural fingerprint (NF) of the molecule which is adapted to the specific task. The architecture has been compared with other GNN models in terms of performance, showing that the proposed approach is very promising.

2) Drug side effect prediction through linear neighborhoods and multiple data source integration

Abstract: predicting drug side effects is a critical task in the drug discovery, which attracts great attentions in both academy and industry. Although lots of machine learning methods have been proposed, great challenges arise with boom of precision medicine. On one hand, many methods are based on the assumption that similar drugs may share same side effects, but measuring the drug-drug similarity appropriately is challenging. One the other hand, multisource data provide diverse information for the analysis of side effects, and should be integrated for the high-accuracy prediction. In this paper, we tackle the side effect prediction problem through linear neighborhoods and multi-source data integration. In the feature space, linear neighborhoods are constructed to extract the drug-drug similarity, namely “linear neighborhood similarity”. By transferring the similarity into the side effect space, known side effect information is propagated through the similarity-based graph. Thus, we propose the linear neighborhood similarity method (LNSM), which utilizes single-source data for the side effect prediction. Further, we extend LNSM to deal with multisource data, and propose two data integration methods: similarity matrix integration method (LNSM-SMI) and cost minimization integration method (LNSM-CMI), which integrate drug substructure data, drug target data, drug transporter data, drug enzyme data, drug pathway data and drug indication data to improve the prediction accuracy. The proposed methods are evaluated on the benchmark datasets. The linear neighborhood similarity method (LNSM) can produce satisfying results on the single-source data. Data integration methods (LNSM-SMI and LNSM-CMI) can effectively integrate multi-source data, and outperform other state-of-the-art side effect prediction methods in the cross validation and independent test. The proposed methods are promising for the drug side effect prediction.

3) Drug Side Effect Analyzer Using Machine Learning

Abstract: People are dependent on medicinal drugs on one way or the other for every simple cause such as headache, cold etc. Every drug has a negative impact on a person's body. Some people are unaware of the side effects of the drugs and they consume it without prescription. Social network platforms such as twitter provide an opportunity for people to express themselves. Using twitter as the source of data, this paper aims to find the side effects of drugs with the help of machine learning algorithms.SVM (Support Vector Machine) algorithm is used for drug related classification with an accuracy of 75%.Sentiment analysis is performed using VADER (Valence Aware Dictionary for sentiment Reasoning) to handle negations, conjunctions and question marks present in the tweets. Keyword Extraction is performed using RAKE (Rapid Automatic Keyword Extraction) to get the side effects.

4) Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources

Abstract: The development of automated approaches employing computational methods using data from publicly available drugs datasets for the prediction of drug side effects has been proposed. This work presents the use of a hybrid machine learning approach to construct side effect classifiers using an appropriate set of data features. The presented approach utilizes the perspective of data analytics to investigate the effect of drug distribution in the feature space, categorize side effects into several intervals, adopt suitable strategies for each interval, and construct data models accordingly. To verify the applicability of the presented method in side effect prediction, a series of experiments were conducted. The results showed that this approach was able to take into account the characteristics of different types of side effects, thereby achieve better predictive performance. Moreover, different feature selection schemes were coupled with the modeling methods to examine the corresponding effects. Additionally, analyses were performed to investigate the task difficulty in terms of data distance and similarity. Examples of visualized networks of associations between drugs and side effects are also discussed to further evaluate the results.

2.EXISTING SYSTMA

drug-drug interaction (DDI) is defined as an association between two drugs where the pharmacological effects of a drug are influenced by another drug. Positive DDIs can usually improve the therapeutic effects of patients, but negative DDIs cause the major cause of adverse drug reactions and even result in the drug withdrawal from the market and the patient death. Therefore, identifying DDIs has become a key component of the drug development and disease treatment. In this study, an existing system, develops a method to predict DDIs based on the integrated similarity and semi-supervised learning (DDI-IS-SL). DDI-IS-SL integrates the drug chemical, biological and phenotype data to calculate the feature similarity of drugs with the cosine similarity method. The Gaussian Interaction Profile kernel similarity of drugs is also calculated based on known DDIs. A semi-supervised learning method (the Regularized Least Squares classifier) is used to calculate the interaction possibility scores of drug-drug pairs. In terms of the 5-fold cross validation, 10-fold cross validation and de novo drug validation, DDI-IS-SL can achieve the better prediction performance than other comparative methods. In addition, the average computation time of DDI-IS-SL is shorter than that of other comparative methods. Finally, case studies further demonstrate the performance of DDI-IS-SL in practical applications.

DISADVANTAGES:

• The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets to detect an accurate Drug Side Effect.

• Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.

• Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

4.PROPOSED SYSTEM

Multi-layer perception (MLP) Our MLP [22] model takes the concatenation of all input vectors and applies a series of fully-connected (FC) layers. Each FC layer is followed by a batch normalization layer [10]. We use ReLU activation [16], and dropout

regularization [27] with a drop probability of 0:2. The sigmoid activation function is applied to the final layer outputs, which yields the ADR prediction probabilities. The loss function is defined as the sum of negative log- probabilities over ADR classes, i.e. the multi-label binary cross-entropy loss (BCE). An illustration of the architecture for CS and GEX features is given in this system. Residual multi-layer perceptron (ResMLP) The residual multi-layer perceptron (ResMLP) architecture is very similar to MLP, except that it uses residual-connections across the fully- connected layers. More specifically, the input of each intermediate layer is element-wise added to its output, before getting processed by the next layer. Such residual connections have been shown to reduce the vanishing gradient problem to a large extend [7]. This effectively allows deeper architectures, therefore, potentially learning more complex and parameter-efficient feature extractors. Multi-modal neural networks (MMNN) The multi-modal neural network approach contains distinct MLP sub-networks where each one extract features from one data modality only. The outputs of these sub-networks are then fused and fed to the classification block. For feature fusion, we consider two strategies: concatenation and summation. While the former one concatenates the domain-specific feature vectors to a larger one, the latter one performs element-wise summation. By definition, for summation based fusion, the domain-specific feature extraction sub-networks have to be designed to produce vectors of equivalent sizes. We refer to the concatenation and summation based MMNN networks as MMNN.Concat and MMNN.Sum, respectively. Multi-task neural network (MTNN) our multitask learning (MTL) based architecture aims to take the side effect groups obtained from the taxonomy of ADReCS into account. For this purpose, the approach defines shared and task-specific MLP sub-network blocks. The shared block takes the concatenation of GEX and CS features as input and outputs a joint embedding. Each task-specific sub-network then converts the joint embedding into a vector of binary prediction scores for a set of inter-related side-effect classes.

ADVANTAGES

The proposed system implemented many ml classifies for testing and training on datasets.

The proposed system developed Convolutional neural networks (CNN) which are known to provide a powerful way of automatically learning complex features in vision tasks to find an accurate accuracy on the datasets.

SYSTEM ARCHITECTURE

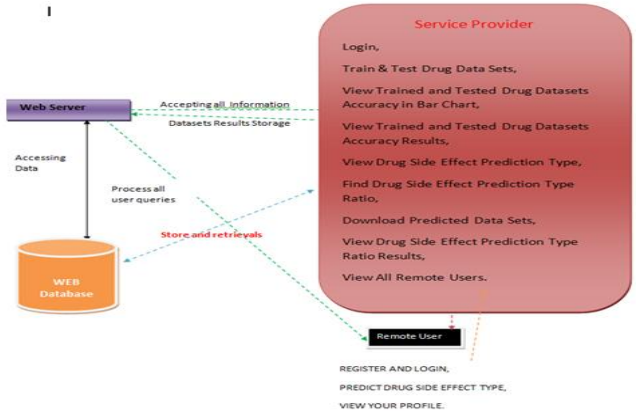

Fig 1: System Architecture

5. UML DIAGRAMS

1. CLASS DIAGRAM

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. It is also known as a structural diagram. Class diagram contains • Classes • Interfaces • Dependency, generalization and association.

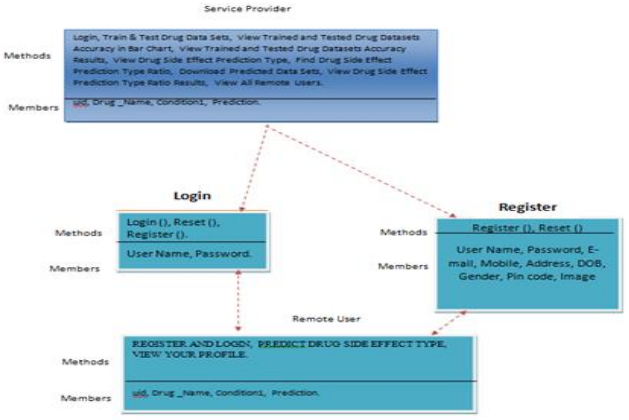

Fig 5.1 shows the class diagram of the project

1. USECASE DIAGRAM:

Use Case Diagrams are used to depict the functionality of a system or a part of a system. They are widely used to illustrate the functional requirements of the system and its interaction with external agents (actors). In brief, the purposes of use case diagrams can be said to be as follows

• Used to gather the requirements of a system.

• Used to get an outside view of a system.

• Identify the external and internal factors influencing the system.

Use case diagrams commonly contains

• Use cases

• Actors

• Dependency, generalization and association relationships.

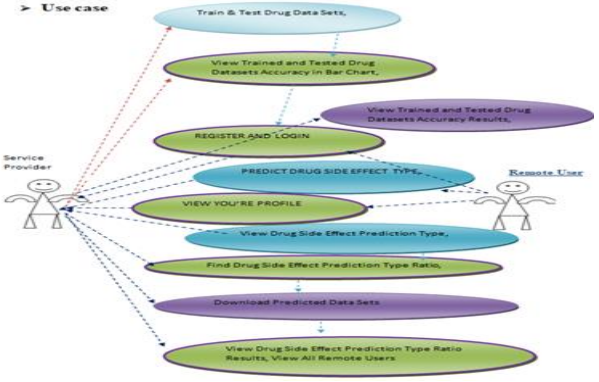

Fig 5.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. Sequence diagrams are used to formalize the behavior of the system and to visualize the communication among objects. These are useful for identifying additional objects that participate in the use cases. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

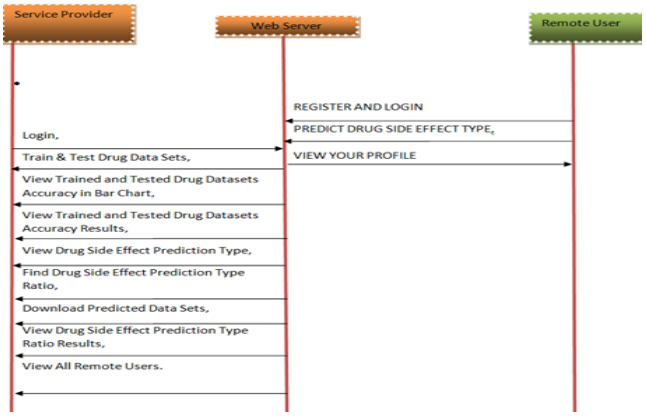

Fig 5.3 Shows the Sequence Diagram

6. RESULTS

Output Screens

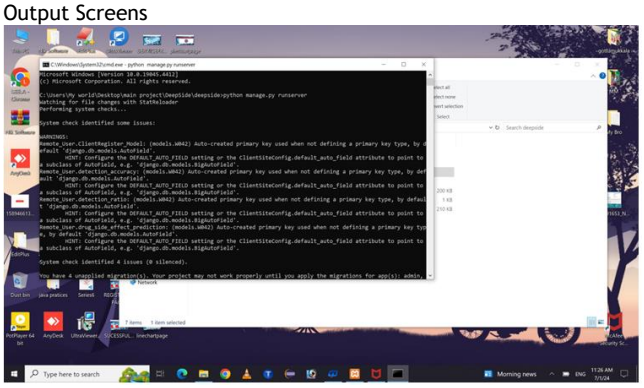

Fig6.1 To Run manage.py File

To run the manage.py file to get the url after that to copy the url and paste into web browser and run to get the home page.

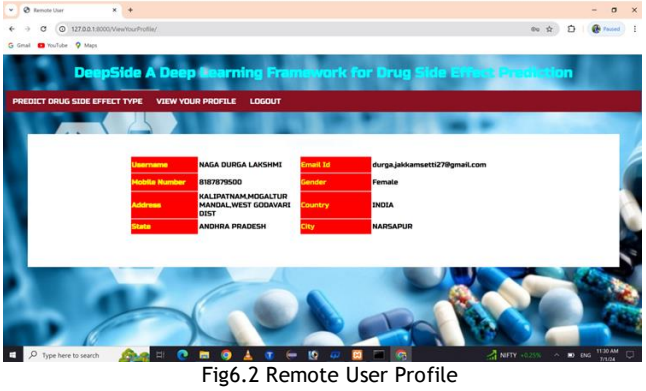

Fig6.2 Remote User Profile

In above screen shows the remote user profile.

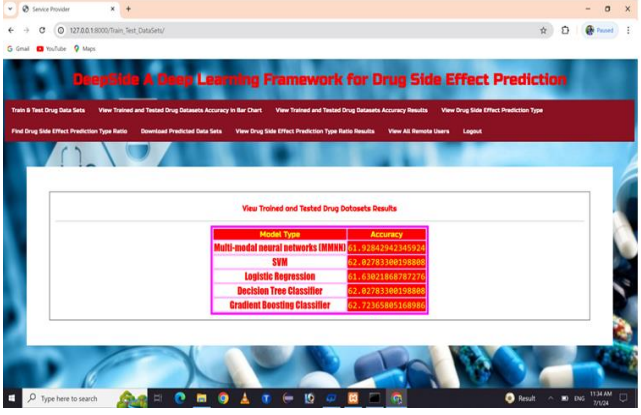

Fig6.3 ML Algorithms Accuracy

To press the upload button it loads the dataset file and then preprocess the dataset after that apply the ml algorithm. The algorithms can train the dataset and produce the accuracy.

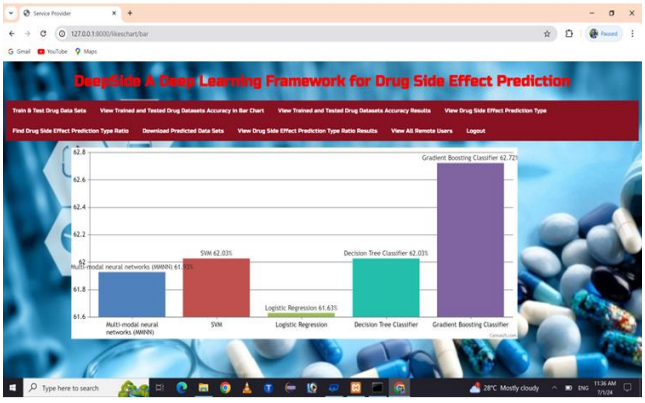

Fig6.4 Bar chart Graph for ML Algorithms

In the above screen shows Algorithm Accuracy in bar chart graph

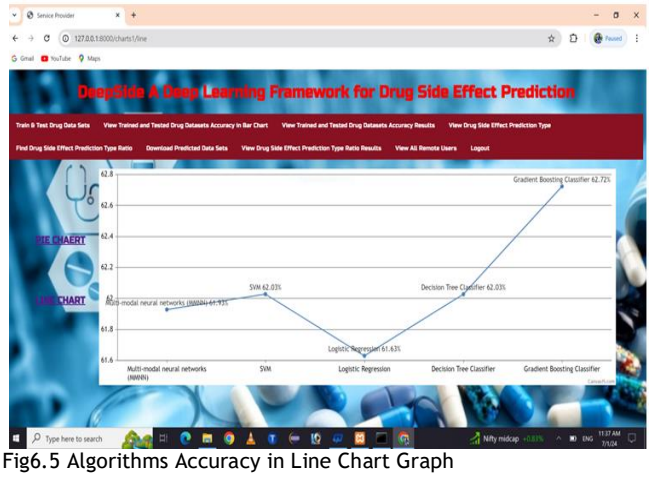

Fig6.5 Algorithms Accuracy in Line Chart Graph

In above screen shows the ml algorithms accuracy in line chart graph.

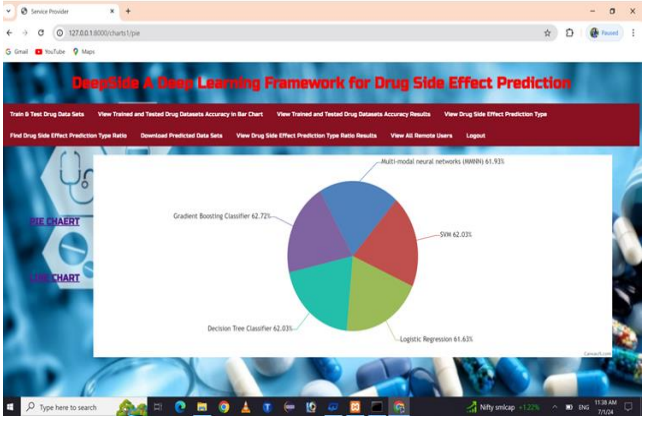

Fig6.6 Algorithms Accuracy in Pie Chart Graph

In above screen shows the ml algorithms accuracy in pie chart graph.

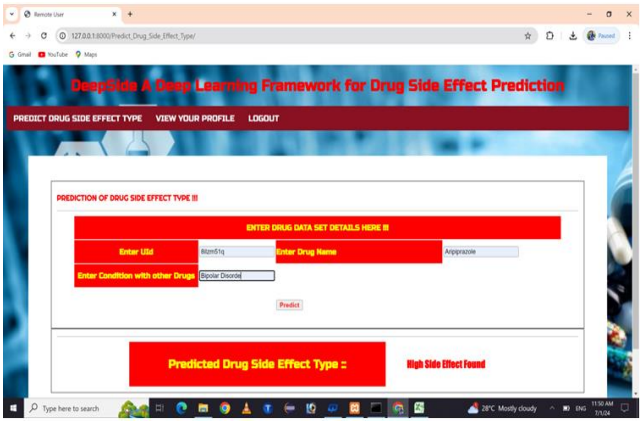

Fig6.7 Prediction of Deep Fake Content

After enter into the remote user login to click on predict button and then get the above page.

CONCLUSION

The pharmaceutical drug development process is a long and demanding process. Unforeseen ADRs that arise at the drug development process can suspend or restart the whole development pipeline. Therefore, the a priori prediction of the side effects of the drug at the design phase is critical. In our Deep Side framework, we use context-related (gene expression) features along with the chemical structure to predict ADRs to account for conditions such as dosing, time interval, and cell line. The proposed MMNN model uses GEX and CS as combined features and achieves better accuracy performance compared to the models that only use the chemical structure (CS) finger- prints. The reported accuracy is noteworthy considering that we are only trying to estimate the condition-independent side effects. Finally, SMILES Conv model outperforms all other approaches by applying convolution on SMILES representation of drug chemical structure.

REFERENCES

- Atias, N., Sharan, R.: An algorithmic framework for predicting side e_ects of drugs. Journal of Computational Biology 18(3), 207{218 (2011)
- Bresso, E., Grisoni, R., Marchetti, G., Karaboga, A.S., Souchet, M., Devignes, M.D., Sma•_l-Tabbone, M.: Integrative relational machine-learning for understanding drug side-e_ect pro_les. BMC bioinformatics 14(1), 207 (2013)
- Cai, M.C., Xu, Q., Pan, Y.J., Pan, W., Ji, N., Li, Y.B., Jin, H.J., Liu, K., Ji, Z.L.: Adrecs: an ontology database for aiding standardization and hierarchical classi_cation of adverse drug reaction terms. Nucleic acids research 43(D1), D907{D913 (2014)
- Dey, S., Luo, H., Fokoue, A., Hu, J., Zhang, P.: Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics 19 (12 2018). https://doi.org/10.1186/s12859-018-2544-0
- Dimitri, G.M., Li_o, P.: Drugclust: A machine learning approach for drugs side effects prediction. Computational Biology and Chemistry 68, 204 { 210 (2017). https://doi.org/https://doi.org/10.1016/j.compbiolchem .2017.03.008
- Groopman, J.E., Itri, L.M.: Chemotherapy-induced anemia in adults: incidence and treatment. Journal of the National Cancer Institute 91(19), 1616{1634 (1999)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
- Huang, L.C., Wu, X., Chen, J.Y.: Predicting adverse side e_ects of drugs. BMC genomics 12(5), S11 (2011)
- Huang, L.C., Wu, X., Chen, J.Y.: Predicting adverse drug reaction pro_les by integrating protein interaction networks with drug structures. Proteomics 13(2), 313{324 (2013)
- Io_e, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
- Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al.: Pubchem substance and compound databases. Nucleic acids research 44(D1), D1202{D1213 (2015)
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classi_cation with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097{1105 (2012)
- Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side e_ects. Nucleic acids research 44(D1), D1075{D1079 (2015)
- Landrum, G., et al.: Rdkit: Open-source cheminformatics (2006)
- Lee, W., Huang, J., Chang, H., Lee, K., Lai, C.: Predicting drug side e_ects using data analytics and the integration of multiple data sources. IEEE Access 5, 20449{20462 (2017).
- https://doi.org/10.1109/ACCESS.2017.2755045
- Li, Y., Yuan, Y.: Convergence analysis of two-layer neural networks with relu activation. In: Advances in Neural Information Processing Systems. pp. 597{607 (2017)
- Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X.w., Matheny, M.E., Xu, H.: Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. Journal of the American Medical Informatics Association 19(e1), e28{e35 (2012)
- Lopes, C.E., Langoski, G., Klein, T., Ferrari, P.C., Farago, P.V.: A simple hplc method for the determination of halcinonide in lipid nano particles: development, validation, encapsulation e_ciency, and in vitro drug permeation. Brazilian Journal of Pharmaceutical Sciences 53(2) (2017)
- Mizutani, S., Pauwels, E., Stoven, V., Goto, S., Yamanishi, Y.: Relating drug{protein interaction network with drug side e_ects. Bioinformatics 28(18), i522{i528 (2012)
- O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R.: Open babel: An open chemical toolbox. Journal of cheminformatics 3(1), 33 (2011)
- Okubo, S., Nakatani, K., Nishiya, K.: Gastrointestinal symptoms associated with enteric-coated sulfasalazine (azul_dine en tablets). Modern rheumatology 12(3), 0226{0229 (2002)
- Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classi_cation. IEEE Transactions on Neural Networks 3(5), 683{697 (Sep 1992). https://doi.org/10.1109/72.159058
- Pauwels, E., Stoven, V., Yamanishi, Y.: Predicting drug side-e_ect pro_les: a chemical fragment-based approach. BMC bioinformatics 12(1), 169 (2011)
- Polatti, F., Viazzo, F., Colleoni, R., Nappi, R.E.: Uterine myoma in postmenopause: a comparison between two therapeutic schedules of hrt. Maturitas 37(1), 27{32 (2000)
- Scheiber, J., Jenkins, J.L., Sukuru, S.C.K., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., et al.: Mapping adverse drug reactions in chemical space. Journal of medicinal chemistry 52(9), 3103{3107 (2009)