IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

Drug Target Interaction Prediction

Ruby Sheikh[1] and Dr. Soumyasri S M[2]
Student, Department of MCA[1]
Associate Professor, Department of MCA[2]
Vidya Vikas Institute of Engineering and Technology, Mysore

Abstract: *Drug–target interaction (DTI) prediction is vital to drug discovery, assisting in the identification of drug- target protein interactions more efficiently than usual methods of experimentation, which are frequently costly and time-consuming. To In order to tackle these issues, this piece presents a web-based application that predicts DTIs and illustrates molecular information. The system takes drug samples and protein sequences as input and outputs molecular formula, predicted activity status with probabilities, and molecular structure visualization. Including ML and DL models, the application enables comparative evaluation of different algorithms, highlighting their strengths and limitations in order to identify most effective approach for DTI prediction. This project's primary purpose is to create reliable models for DL and ML that can precisely determine whether a medication and protein combination is active or inert. Because of this, it is feasible to forecast drug-target interactions with accuracy. Data on drug-protein interactions was acquired from ChEMBL then was pre- processed and represented using SMILES, molecular fingerprints, and ProtBERT embeddings. Multiple ML and DL models (RF, SVM, KNN, Decision Trees, Logistic Regression, Naïve Bayes, CNN, RNN) were trained, and their performance was reviewed using accuracy, precision, recall, and ROC-AUC to determine the most effective approach. Logistic Regression achieved highest performance with the highest accuracy of 96% which outperformed several deep learning approaches on this dataset, however deep learning retains advantage of automatic feature learning and may surpass machine learning when larger and more dataset available. This dual evaluation confirms our hypothesis on both sides in a much more robust way. This comparison provides hands-on experience on benefits and drawbacks of the two approaches.*

Keywords: DTI, Machine Learning, Deep Learning, Molecular Fingerprints Protein Sequence Embedding Feature Extraction Random Forest(RF) Recurrent Neural Network(RNN) Convolutional Neural Network (CNN).

I. INTRODUCTION

Understanding how drugs bind their protein targets is a crucial aspect of drug discovery. It is the kinds of interactions here that explain whether and how a drug can work in the body. If a drug does stick to that protein on the right, it may appear very much like a promising therapy and if they don't, then not only is the drug likely to be a dud — it could also introduce side effects. Therefore, the costs of not understanding these opposing events are cost effective in terms of then developing a drug when one side inhibits while the other stimulates. In history, drug-target interactions was detected in wet-lab. These strategies to accomplish include high-throughput screening, affinity assays and molecular docking. While the accuracy of these methods are only too good, there are serious limitations to them. They're also slow and costly — not great for testing the millions (or more) of potential drug/protein pairings you might want to explore. To overcome these issues, computational approaches have gained traction. Rather than testing everything in the lab, therefore, researchers can instead tap algorithms to predict what intervenes in a given biological circuit what interactions are likely. It's a way to focus the list of candidates before experimental tests. Of the computational approaches employed in this regard, machine learning (ML) and deep learning (DL) have provided significant results lately. Machine learning algorithms, such as RF, SVM and LR have been applied extensively in DTI prediction. These models use features describing drugs and proteins. For instance, drugs can be expressed as molecular fingerprints and proteins in terms of their amino acid sequence. However, conventional ML models usually depend on manually-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

designed features. These characteristics are occasionally too simplistic to describe those complex relations among drugs and proteins. Deep learning offers more flexibility. It can learn patterns from raw or inadequately processed data. CNNs (Convolutional Neural Networks) are effective for processing of spatial data, such as molecular structures. It also uses RNNs (Recurrent Neural Networks) such as LSTM model, which is more efficient for sequence data, like proteins. More recently, transformer-based models like ProtBERT have been used to create rich embeddings of protein sequences. Such approaches avoid the necessity of hand design features. In this work we contrast ML and DL approaches in developing A more precise model to forecast interactions between drugs and their targets. We source our data from publicly available resources. Drugs are transformed to molecular fingerprints using some tool such as RDKit. Models are used to encode protein sequences such as ProtBERT. We actually try different algorithms; traditional ML and DL-based methods to find out which one is best. An ideal prediction system is expected to be able to predict potential drug-target associations accurately. This might save scientists money by decreasing the need for expensive laboratory experiments, and also speed drug discovery itself in its earliest stages.

II. RELATED WORK

In the prediction field Drug-Target Interaction (DTI) novel research is being done on approaches that now utilize a mix of traditional ML and Deep Learning (DL) which has fastened the evolution process for early-stage drug discovery. Traditional ML models like Logistic Regression, SVM and Random Forest have found heavy usage in practice because of their interpretability, the need for a small-time complexity, and relatively good results on small datasets. In contrast, DL models like for instance, convolutional and RNN, including LSTM networks automatically learn complex patterns from molecular and protein sequence data for deeper insights on large datasets. The public datasets, including ChEMBL, BindingDB and DrugBank, are popular benchmarks for DTI studies (which can be considered as molecular descriptors calculated by RDKit and protein embedding produced by ProtBERT). While DL models are good in capturing non- linear relations, it has been shown that ML models can beat them in the returned or small data scenario. Based on this finding, we integrate ML and DL methods in ChEMBL data set to develop predictive mode with comparative performance for DTI.

SYSTEM DESIGN AND IMPLEMENTATION

- Drug data input – Users insert chemical information (e.g. SMILES strings or molecular fingerprints).
- Protein Data Input – They introduce the sequences of target proteins.
- Data Preparation: Inputs are cleaned, standardised and transformed in a format to feed the model.
- Feature generation – drugs and proteins are encoded to generate machine-interpretable features.
- Model Designing – Pattern of interactions are obtained by using machine learning or deep learning techniques.
- Interaction Prediction–The trained model predicts the interaction probability for a given drug-protein pair.
- Results Display – Users obtain predicted outcomes along with performance indicators

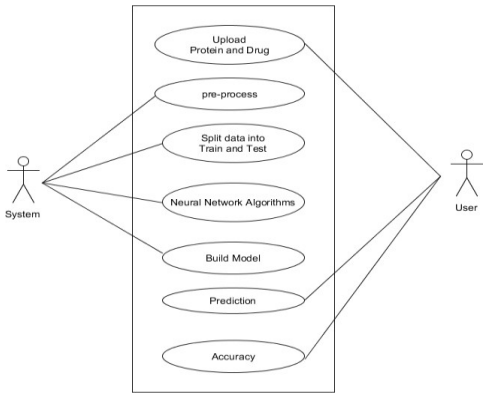

Fig: Use Case Diagram.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

III. METHODOLOGY

Data Collection Layer: Datasets are sourced primarily from ChEMBL, providing molecular structures in SMILES format and sequences of proteins in FASTA format. While other datasets were evaluated during prototyping, ChEMBL serves as the main source for the implemented system.Pre-processing Layer: Here raw data is converted & shaped into machine readable formats that can be processed immediately. Drug molecules are featured using RDKit to generate molecular descriptors and fingerprints, and protein sequences are embedded with ProtBERT for numerical vectors in high dimensions.

Feature Engineering: The drug and protein features can be processed into unified vectors that are dedicated to ML as well as DL methods. Dimensional reduction methods including PCA and t-SNE are used to accelerate computation. Model Training Layer: Besides the deep learning- based CNN and RNN models, the framework fuses traditional methods, machine learning such as Random Forest (RF), SVM, KNN, Decision Trees, Logistic Regression (LR) and Gradient Boosting.

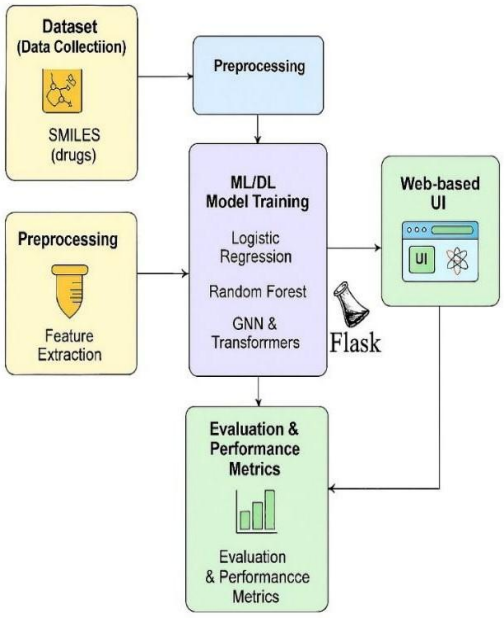

Fig 1: Speech Recognition

Evaluator layer: The model is tested, and obtain the ROC- AUC, F1-score, recall, accuracy and precision of each scenarios.Cross-validation actually makes your models generalise well on new data.

Deploy Layer: The most precise model is deployed as the Flask API to obtain an easy-to-use web interface obtaining user- friendly drug–protein interaction prediction in real time.

IV. RESULT AND DISCUSSIONS

In this study, we compare classical ML and DL models for the DTI prediction problem given a set of labeled drug-protein pairs. Model recall, F1-score and ROC-AUC figures. Among the classical models, Logistic Regression does largely better with an overall 96% of score on all metrics compared to the others and Naïve Bayes (with an accuracy from 93 to 94%) – although it handles very well minority classes. SVM and CNN reached around 85%, indicating conservative predictions, while KNN and Decision Tree showed moderate performance (accuracy 76–80%). Gradient Boosting and RNN (LSTM) experienced the largest performance drop, particularly RNN (accuracy 48%, F1-score 0.31). The models with 82–83% sensitivity implemented more slow development, but seemed to scale and learn good features from raw inputs. CNN learnt molecular features of interest (85% accuracy) and RNN-LSTM utilized protein sequences (49%). The learning curves showed stable convergence with little overfitting. The classical models worked well with the predefined engineered features, and then the DL model provided a means to derive feature from raw

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

input. The models were integrated inside a Flask web application, which allows for real time predictions of SMILES and FASTA inputs with confidence scores in less than 2 seconds on an average, signatory to practicality of the proposed model in DTI screening.

V. CONCLUSION

This review is focused on application of ML+DL (also its older sister, ML) for DTI prediction. The model was generally fitted in order to generate an effective approach of exploring the potential medicinal molecules that could bind well with some specific target protein sequence. Through extensive experiments and analyses, we provided evidence that both classical ML models (i.e. Random Forests, Gradient Boost) and the modern DL model (RNN,CNN) could be successfully used for accurate prediction of the DTI.

This study highlights the value of ML models in the initial stages of drug discovery, especially when comes to using protein sequences to predict drug–target interactions. Drug screening procedures can be expedited, experimental expenses decreased, and therapeutic research made more focused and effective by utilizing these models between Deep learning and DL models. Logistic Regression achieved highest performance with the highest
accuracy value of 96% which outperformed several deep learning approaches on this dataset, however
deep learning retains advantage of automatic feature learning and may surpass machine learning when larger and more dataset available. This dual evaluation strengthens our understanding of both with larger datasets. This comparative evaluation provides the valuable insights into strengths and weakness of both approaches.

REFERENCES

[1] S. M. Moussa, M. R. Barkat, and N. L. Badr, "Drug-target interaction prediction using machine learning," in 2021 10th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), Cairo, Egypt, 2021, pp. 480–485, doi: 10.1109/ICICIS52592.2021.9694127.

[2] X. Zhang, Q. Ye, and X. Lin, "Drug-target interaction prediction via multiple output deep learning," in 2020 IEEE Int. Conf. Bioinf. Biomed. (BIBM), Seoul, South Korea, 2020, pp. 507–510, doi: 10.1109/BIBM49941. 2020.9313488

[3] D. Sathan and S. Baichoo, "Drug-target interaction prediction using variational quantum classifier," in 2024 Int. Conf. Next Gener. Comput. Appl. (NextComp), Mauritius, 2024, pp. 1–7, doi: 10.1109/NextComp63004.2024.10779674.

[4] P. Razzaghi, K. Abbasi, A. Poso, and S. G. Ara, "Deep learning in drug-target interaction prediction: Current and future perspective," Curr. Med. Chem., 2020, doi: 10.2174/0929867327666200907141016.

[5] Z. Liao, X. Huang, H. Mamitsuka, and S. Zhu, "Drug3D-DTI: Improved drug-target interaction prediction by incorporating spatial information of small molecules," in 2021 IEEE Int. Conf. Bioinf. Biomed. (BIBM), Houston, TX, USA, 2021, pp. 340–347, doi: 10.1109/BIBM52615.2021.9669707.

[6] L. Xie, Z. Zhang, S. He, X. Bo, and X. Song, "Drug–target interaction prediction with a deep-learning-based model," IEEE Int. Conf. Bioinf. Biomed. (BIBM), pp. 469–476; Kansas City, MO, USA, 2017 doi: 10.1109/BIBM.2017.8217693.

[7] H. Yang, "MINDG: A drug–target interaction prediction method based on an integrated learning algorithm," Bioinformatics, vol. 40, no. 4, btae147, 2024, doi: 10.1093/bioinformatics/btae147.

[8] W. Yin, J. Wang, G. Zhang, H. Luo, W. Liang, J. Luo, and C. Yan, "Drug–drug interactions prediction based on deep learning and knowledge graph: A review," iScience, vol. 27, no. 3, 109148, 2024, doi: 10.1016/j.isci.2024.109148.

[9] K. Faez, H. Abbasi Mesrabadi, and J. Pirgazi, "Drug–target interaction prediction based on protein features, using wrapper feature selection," Sci. Rep., vol. 13, 3594, 2023, doi: 10.1038/s41598-023-30026-y.

[10] N. H. A. Hassain Malim, H. Ismail, S. Z. M. Zobir, and H. A. Wahab,"Comparative studies on drug–target interaction prediction using machine learning and deep learning methods with different molecular descriptors," in 2021 Int. Conf. Women Data Sci. Taif Univ. (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1–6, doi: 10.1109/WiDSTaif52235.2021.94301.