International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
Volume 11, Issue-2 March-April- 2024
Email: editor@ijermt.org www.ijermt.org

MACHINE LEARNING MODELS FOR PREDICTING DRUG TARGET INTERACTIONS

Ms Shikha Rajput

Assistant Professor,CCS University,Meerut

Dr Yashwant Rai

Assistant Professor ,CCS University,Meerut

ABSTRACT

Within the realm of drug development, machine learning (ML) has emerged as a crucial tool, particularly for the purpose of predicting Drug-Target Interactions (DTIs). The purpose of this research is to investigate the use of machine learning models in DTI prediction, with a particular focus on the role that these models play in accelerating the identification of new treatments. The ability of machine learning models to efficiently understand complex interactions between medications and target proteins is made possible by the utilization of enormous biological and chemical information as well as the application of advanced algorithms. Data collection, feature representation, model selection, evaluation, and interpretation are some of the stages that are essential in the process of constructing machine learning models for DTI prediction. A number of challenges, including the quantity of the dataset, the interpretability of the model, and the validation of the experiment, continue to exist, but they are being solved through collaboration across disciplines. Despite these obstacles, it seems that machine learning will have a bright future in the field of DTI prediction. It will provide potential to speed up the process of drug discovery and to reinvent therapeutic development.

Keywords: *Machine ,learning, Drug*

INTRODUCTION

Over the course of the past several years, there has been a substantial increase in the utilization of machine learning (ML) strategies in the exploration and development of pharmaceuticals. In the process of drug development, one of the many obstacles that must be overcome is the prediction of interactions between medicines and the proteins that they are intended to have an effect on. When it comes to understanding the mechanism of action of medications and locating prospective therapeutic targets, this interaction, which is sometimes referred to as Drug-Target Interaction (DTI), plays a crucial role. In the past, the process of identifying DTIs depended mainly on experimental procedures, which are not only expensive and time-consuming, but also frequently restricted in their breadth. On the other hand, researchers have turned to predictive modeling approaches in order to speed up the process of discovering prospective DTIs. This is due to the exponential expansion of biological and chemical data, as well as breakthroughs in machine learning algorithms and processing capacity.

Through the process of discovering patterns and associations from enormous amounts of biological and chemical data, machine learning models provide a strong method for predicting different types of DTIs. Molecular structure, physicochemical characteristics, biological pathways, and sequence information are some of the features that are utilized by these models. These models also make use of other data that are retrieved from medications and target proteins. Through the utilization of complicated algorithms and the incorporation of a wide variety of data sources, machine learning models are able to efficiently capture the intricate interactions that occur between medications and targets.

Applications

Figure 1 illustrates the drugs and targets that are currently being used for drug–target interaction prediction. Studies that have already been conducted on the topic of drug–target interaction prediction have demonstrated that making use of various calculation or optimization approaches during the stages of data set collection, feature extraction and processing, and task algorithm selection can result in the construction of models that have a high level of performance.

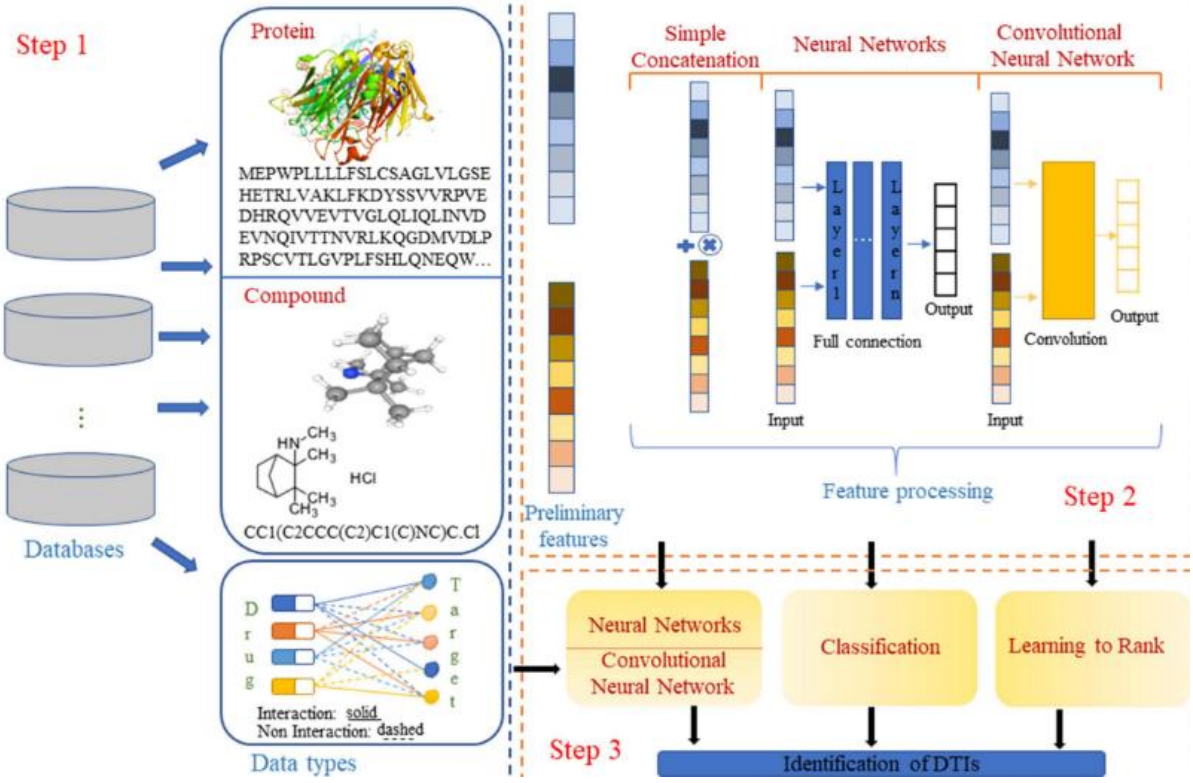

Figure 1. **Methods for foretelling the interactions between drugs and their targets. Sourced from PubChem are the drug's two- and three-dimensional structural diagrams..**

First, gathering the dataset. Excessive experimental cycles and skewed results might be the consequence of redundant data, imbalanced categories, and samples that do not accurately reflect the population. These issues have been mitigated or eliminated from the model-building process by employing alternative data collecting techniques. Among other things, we addressed the issue of data imbalance by collecting negative cases by random selection. moreover employed a random selection procedure to glean negative cases; this process

was repeated five times to mitigate the influence of the unverified negative samples. Pdti-EssB managed to resolve the issue of data imbalance by employing under-sampling clustering and random under-sampling.

At present, the majority of target molecules are proteins. Out of all the target molecules, 44% are kinases, G protein-coupled receptors (GPCRs), ion channels, and nuclear receptors. Of the medications being researched, 70% are specifically designed to target these four groups of proteins. Databases that have been created that include the drug-protein interactions involving these four proteins have seen extensive use. The relevant Using these datasets, the majority of computer techniques have narrowed their attention to just investigating the possibility of a drug's interaction with a certain protein, a practice known as binary classification. Some research has looked into drug-target affinity as a way to speed up the process and cut costs even more. An important feature that establishes the intensity of the connection between the target and the small molecule medicine is the drug-target affinity. The KIBA and Kinase databases are the most popular choices for drug-target affinity prediction.

(2) Processing and extraction of features. For high-performance model building, precise and thorough numerical representations of drug and target biological or chemical functional information are crucial. There are several angles from which to obtain medication and target features. To illustrate, iGPCR-Drug derives drug characteristics from discrete Fourier transforms of drug molecular fingerprints. GPCR characteristics based on amino acid compositions that are not real. DrugE-Rank is able to extract target properties based on amino acid composition, transformation, and distribution, and it also displays drug attributes according to generic descriptors. Using wavelet transform on drug molecular fingerprints, TargetGDrug extracts drug characteristics. Evolutionary information determines GPCR characteristics. derived drug characteristics based on generic descriptors and retrieved protein features using the distance-based top-n-gram technique. Chemical databases often employ text as their data storage format, and a lot of cheminformatics programs adhere to the simplified molecular input line entry specification (SMILES) format. Complex chemical characteristics may be predicted using the structural information encoded by each SMILES string. Many machine learning methods can extract molecular attributes of compounds based on SMILES strings. Molecular feature extraction has recently made use of recurrent neural networks and convolutional neural networks (CNNs). employed convolutional neural networks (CNNs) to extract molecular characteristics after transforming SMILES strings into two-dimensional matrices. used recurrent neural networks to extract features from SMILES and natural language processing to handle molecular strings.

The accuracy of the experiment result and the length of the experimental period are both affected by the existence of faulty or redundant characteristics. We anticipate feature sets of information that are both comprehensive and low-dimensional. Consequently, related rearch have been subjected to a number of feature processing approaches. By reducing the dimensionality of medication and target data using principal component analysis (PCA), for instance, the noise between features was reduced. formed feature vectors of drug-target couples by combining 881 drug substructures and 876 target Pfam domain structures using tensor product. To create a novel, low-dimensional representation of drug and protein properties, MFDR employed autoencoders, which are the building blocks of deep networks. By convolutionally weaving amino acid subsequences of varying lengths, DeepConv-DT was able to extract local amino acid residue information from raw protein sequences using convolutional neural networks (CNNs).

Three, algorithms for task selection. Predicting drug-target interactions has made use of a number of task algorithms, including deep learning, learning to rank, and classification algorithms.

A variety of classification techniques have been utilized in the majority of the existing research, which views drug-target interaction prediction as a binary job. To forecast drug-target connections, for instance, a bipartite local model (BLM) with a support vector machine (SVM) kernel was suggested. By combining Lasso for feature extraction with random forest for classification, LRF-DTI is able to predict drug-target interactions. used a classifier based on a distance learning algorithm. To determine if medications and targets are dockable, pred-binding employed support vector machines and random forests to categorize characteristics taken from molecular structures and protein sequences.

Predicting drug-target interactions is similar to a ranking problem. A more efficient and cost-effective drug development process can be achieved by investigating the potency of drug-target interactions. investigated the application of six learning-to-rank algorithms—Prank, RankNet, RankBoost, SVMRank, AdaRank, and ListNet—to virtual drug screening. The results demonstrated that learning-to-rank is a powerful computational strategy, particularly for its innovative applications in cross-target virtual screening and heterogeneous data integration. To enhance the efficacy of drug-target interaction prediction, DrugE-Rank utilized variables such as protein amino acid composition, information about transformation and distribution, information about compounds, and the output of six classifiers as inputs into a learning-to-ranking system.

Predicting drug-target interactions is another area that has shown success with neural networks. used neural network data and entropy information from drug-protein complexes to forecast affinity values for therapeutic targets. A unique method utilized in this study is the modeling of protein sequences and compound 1D representations with convolutional neural networks (CNNs). DeepDTA suggested a model based on deep learning that relied solely on sequence information of both medicines and targets. Using a graph convolutional network to learn the drug-target binding affinity, graph DTA centered on the idea that molecules are naturally created by chemical bonding of atoms.

Method

The purpose of this study is to provide a method for identifying DTIs that is based on machine learning. In the initial step of this approach, various characteristics are extracted from the sequence of proteins, and then the feature vector of proteins is constructed. A fingerprint is then derived from the structure of the medicine once it has been analyzed. The IWSSR approach is then used to pick the features once they have been merged. This is because the features have a high dimension, which makes it difficult to select them individually. After that, the rotating forest model is taught to recognize interactions, which finishes the process. In Figure 2, the flowchart of the suggested approach is displayed. In the following, you will find the specifics of each stage.

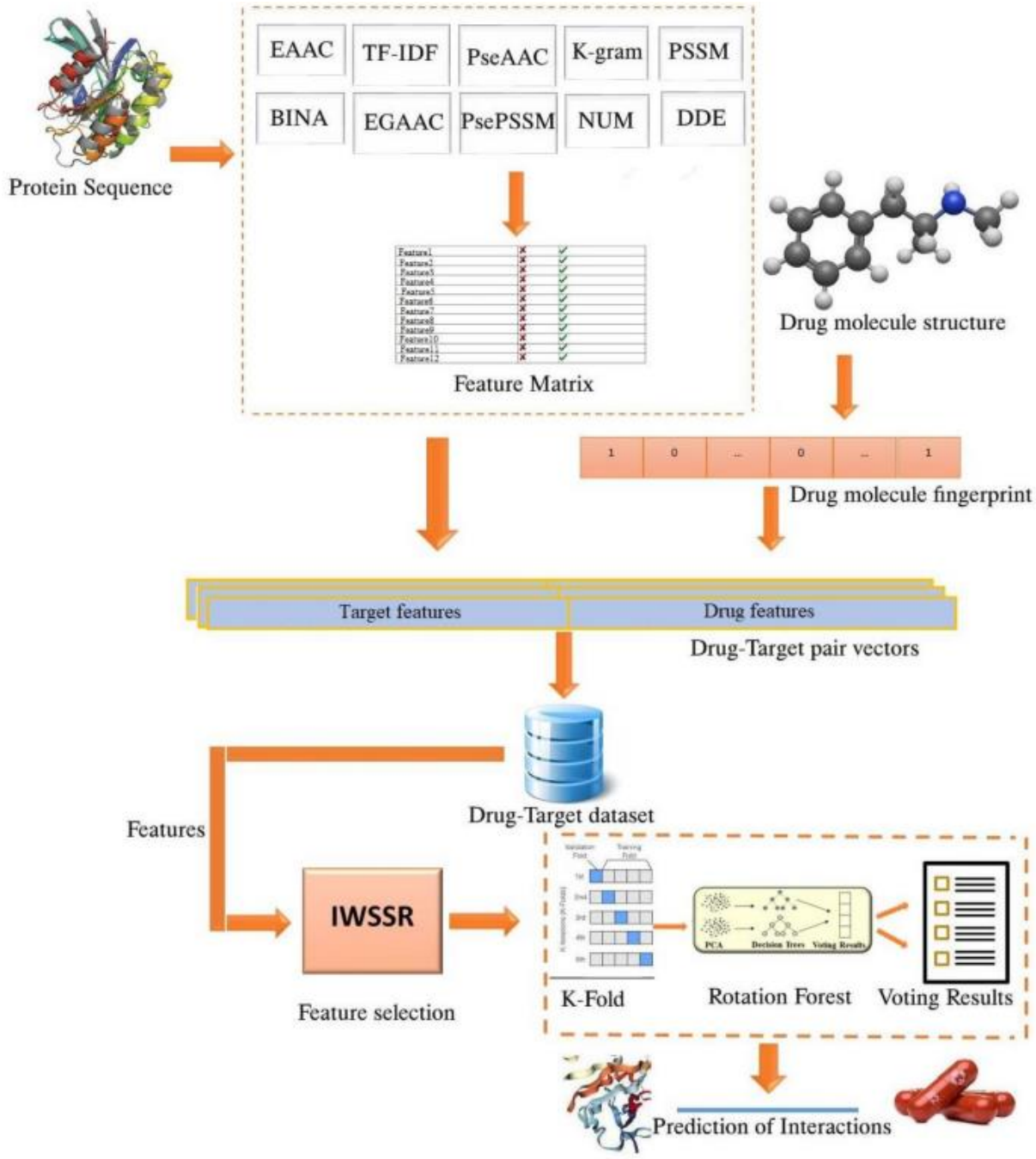

Figure 2 General steps of the proposed method.

Feature extraction

At this stage, a feature extraction method is utilized to return the information of each sequence to a numeric vector. This is done in order to complete the process. This step is one of the most significant phases in the classification phase, and it will have a direct impact on the outcomes of the model prediction. In light of the fact that this investigation makes use of two different inputs, namely pharmaceuticals and proteins, the process of feature extraction may be broken down into two distinct categories: feature extraction from drugs and feature extraction from proteins.

Feature extraction of drugs

It has been demonstrated by researchers that molecular fingerprints are capable of describing the structure of a medicine. Through the process of dividing the molecular structure of pharmaceuticals into a number of different parts, the fingerprint of structural connections reveals that drugs are the vectors of Boolean substructure.

This ensures that the structural information of the entire medicine is maintained, despite the fact that each molecule is broken up into its component components. When it comes to the description and screening process, these descriptors have the ability to reduce the likelihood of information failure and interactions that are not wise. In specifically, a dictionary that has already been predefined and has all of the infrastructures that correlate to the fragments of the pharmacological molecule. The position of a fragment on the user's device is deemed to be "one" if it is found in the dictionary; otherwise, it is considered to be "zero" if it is not inside the dictionary. In the form of binary fingerprint vectors, the database of the whole fingerprint facilitates the creation of an efficient method for the description of the production of the drug molecules. A map of the chemical formation that was obtained from the PubChem system is utilized in this particular piece of writing. A total of 881 molecular infrastructures are included in this system. For this reason, the 881-dimensional binary vector format has been utilized for the descriptors of the structure of drug molecular characteristics.

Feature extraction of proteins

The extraction of key characteristics from protein sequences is part of the process of detecting DTIs, and it is one of the most critical steps. Protein sequences have been analyzed in this work in order to extract a variety of characteristics for this aim. EAAC, EGAAC, DDE, TF-IDF, k-gram, BINA, PSSM, NUM, PsePSSM, and PseAAC are some of the properties that are included in this category. Both the description and the method of feature extraction for each are described in the following paragraphs:

Enhanced amino acid composition (EAAC)

This approach was suggested by Chen and his colleagues. In this approach, information about the sequence of the protein is retrieved, and then the frequency information of the amino acids is determined based on that information. For the purpose of calculating this approach, the following equation is necessary:

$$g(m,n) = \frac{H(m,n)}{H(n)}, m \in \{1,2,\ldots,21\}, n \in \{W_1, W_2, \cdots, W_L\} \quad (1)$$

In this connection, the letter m represents the amino acids, the letter n represents the numerous windows of varying sizes, the letter H(m,n) represents the number of amino acids of type m, and the letter H(n) represents the longitude of the window.

Enhanced grouped amino acid composition (EGAAC)

Protein sequences are transformed into numerical vectors using this technique, which is based on the characteristics of the sequences. One of the most significant feature elicitation algorithms is this method,

which is used in the field of bioinformatics research, namely for the prediction of malonation sites and other related topics. There are twenty distinct types of amino acids, and they are categorised into five groups based on five physical and chemical characteristics (physicochemical): The aliphatic group is comprised of GAVLMI amino acids, the aromatic group is comprised of GFYW amino acids, the positively charged group is comprised of KRH amino acids, the negatively charged group is comprised of DE amino acids, and the uncharged group is comprised of STCPNQ amino acids. It is recommended that the following equation be used for the computation of EGAAC, and this recommendation is based on the basis of this gathering:

$$G(g,n)=\frac{H(g,n)}{H(n)}, g\in\{g_1,g_2,g_3,g_4,g_5\}, n\in\{W_1,W_2,\ldots,W_L\} \quad (2)$$

In this equation, the symbol H(g,n) represents the number of amino acids that belong to group g in window n, while the symbol H(n) is equivalent to the longitude of window n. Within the scope of this investigation, the window size is regarded as L-5, where L represents the length of the protein sequence.

Dipeptide deviation from the expected mean (DDE)

In the field of feature extraction based on amino acid composition, the Dipeptide Deviation technique from the anticipated mean (DDE) has been presented and developed in order to differentiate epitopes of a cell from non-epitopes by utilising this feature extraction approach. This method has been researched in the field of feature extraction. In order to accomplish this, the dipeptide composition of a protein sequence (DC sequence) is initially determined by the following formula:

$$DC(m,n)=\frac{H_{mm}}{H-1} \quad m,n\in\{A,C,D,\ldots,Y\} \quad (3)$$

The results and discussion

In this section, we will present the empirical results of our proposed prediction model for DTIs that was implemented on two datasets. These datasets include protein sequences, drug SMILES (1D raw data), and features data. Scikit-learn, ensemble package, kares library, tensorflow library, and XGBoost package are the tools that are utilised in the application of each technique. The Python programming language is version 3.6.

The accuracy, mean square error, mean squared error, and f-score that were achieved by various methods are reported in Table 1, which contains the results. Using the benchmark dataset, LightBoost and ExtraTree ensemble learning were able to reach the highest accuracy score value of 0.98, while RF was able to acquire the second best value of 0.97. ExtraTree ensemble learning achieved the greatest precision score value of 0.966 for the DrugBank dataset, while Random Forest achieved the second highest value of 0.96. Both of these scores are considered to be the best. In addition, the ExtraTree method offers the highest F1-score for this particular forecast.

Table 1 Accuracy, Mean Square Error, MCC Score, and F1-score metrics used to evaluate the performance of deep, machine, and ensemble methods.

Algorithm	Dataset	Accuracy Score	Mean Square Error	MCC Score	F1-score
ANN	DrugBank	0.9277	0.072	0.848	0.88
	Benchmark	*0.9718*	*0.024*	*0.95*	*0.953*
DBN	DrugBank	0.917	0.056	0.89	0.885
	Benchmark	0.94	0.02	0.95	0.92
Random Forest(RF)	DrugBank	*0.947*	*0.0528*	*0.887*	*0.927*
	Benchmark	0.9744	0.0257	0.945	0.96
SVM	DrugBank	0.93	0.07	0.85	0.915
	Benchmark	0.96	0.039	0.917	0.948
LightBoost	DrugBank	0.938	0.0197	0.958	0.918
	Benchmark	*0.98*	*0.0613*	*0.869*	*0.974*
XGBoost	DrugBank	0.913	0.087	0.814	0.88
	Benchmark	0.97	0.029	0.938	0.96
ExtraTree	DrugBank	0.94	0.056	0.88	0.915
	Benchmark	*0.98*	*0.016*	*0.965*	*0.978*

Table 2 presents a comparison of the various approaches with regard to the amount of time required for the model training process. According to the data presented in the table, Random Forest is the technique that achieves the best results, with a running time of 1.78 seconds and 1.28 seconds when applied to the two datasets. Additionally, ExtraTree ensemble approaches also get a decent result with a training time of 1.79 seconds. Obtaining in the DBN approach by 14103.78 seconds and 7821.48 seconds for the two separate datasets is the execution time that is considered to be the worst case scenario. CNN's time is significantly longer than that of the DBN algorithm, whereas CNN's time is twice as long.

Table 2 Deep, machine, and ensemble methods' Time-based outputs.

Algorithm	**Dataset**	**Time in seconds**
ANN	DrugBank	518.8

	benchmark	501.5
DBN	DrugBank	*14103.78*
	benchmark	*7821.48*
CNN	DrugBank	28080
	benchmark	15642
Random Forest(RF)	DrugBank	1.78
	benchmark	1.28
SVM	DrugBank	184.6
	benchmark	53.12
LightBoost	DrugBank	10.1
	benchmark	12.31
XGBoost	DrugBank	90.1
	benchmark	52.14
ExtraTree	DrugBank	*1.79*
	Benchmark	*0.796*

In order to offer a more accurate visual interpretation for Drug Target Interactions prediction, the area under the curve (AUC) is produced for each model based on the ROC curve. This is done in order to represent the quality of the job. As shown in Figure 3, the ROC curve and the value of the area under the curve (AUC) for each of the learning techniques are displayed. The random forest and artificial neural network (ANN) methods predict the maximum value in the area under the curve (AUC) for DrugBank datasets, which is 0.937. The additional tree technique predicts the maximum value in the AUC, which is 0.982, for the DrugBank data set in the benchmark data set.

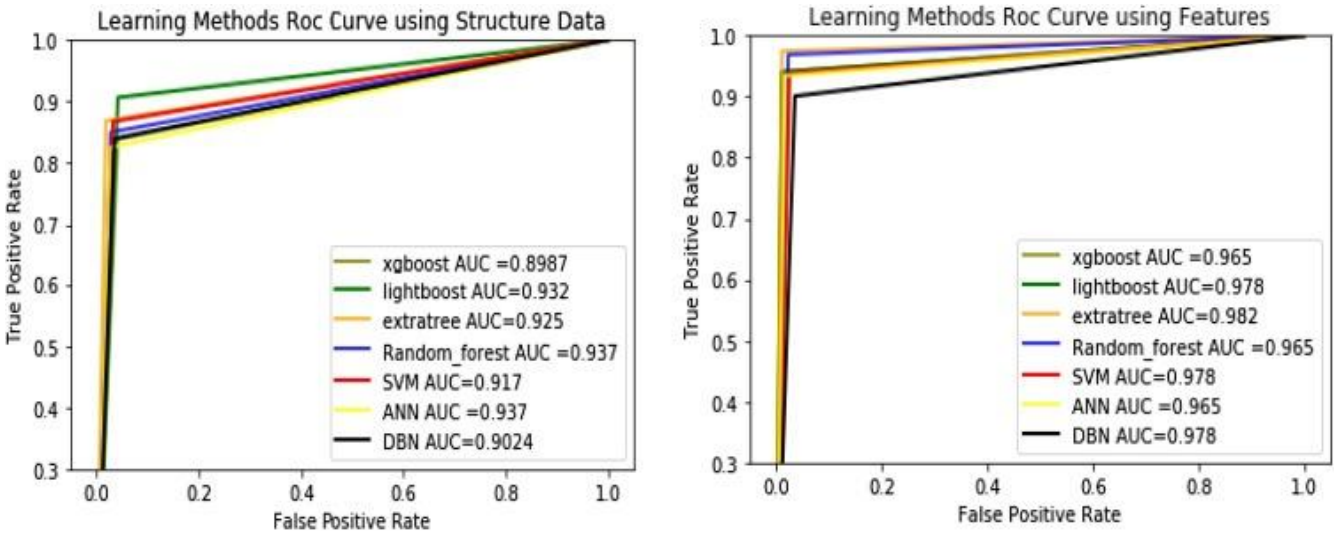

Fig. 3 In the DrugBank data set, the ANN and random forest methods predicted a maximum value of AUC = 0.937, while the extra tree method predicted a maximum value of AUC = 0.982, according to the results of the ROC curve and the value of the area under the curve (AUC) for the learning methods.

As shown in Figure 4, the Precision-Recall (PR) Curve is a straightforward graph that displays the values of Precision along the y-axis and the values of Recall along the x-axis simultaneously. Note : Precision is frequently referred to as the Positive Predictive Value (PPV), which is an essential distinction to make. Sensitivity, Hit Rate, and True Positive Rate (TPR) are all names that may be used to refer to recall on Davis (2006). The Random Forest approach is the one that provides the best precision recall curve when it comes to sequence data, while the DBN method is the one that stands out when it comes to features data.

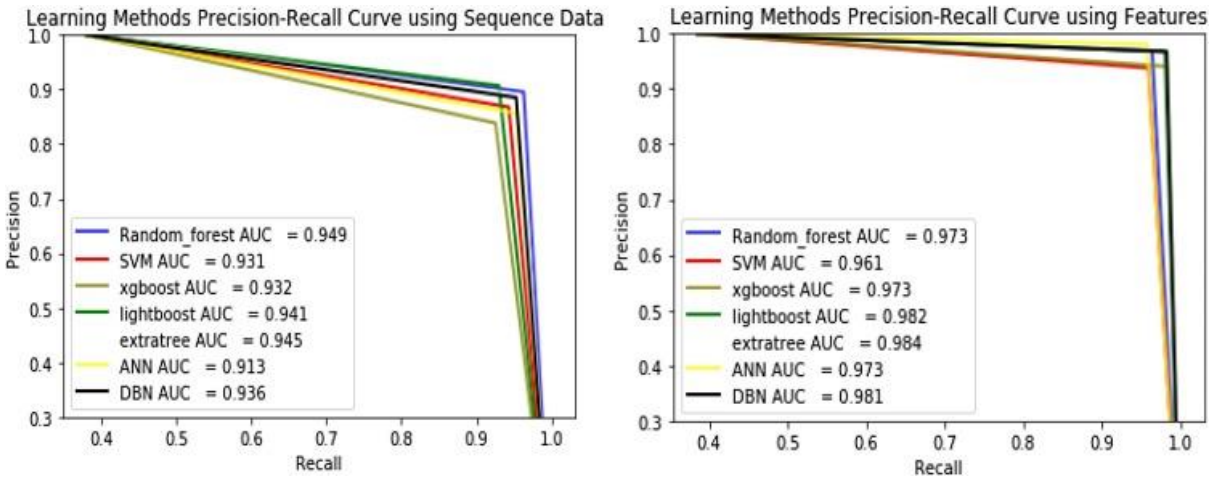

Fig. 4 displays the data set's recall and accuracy curve for features and sequences. A distinct cutoff is shown by the tradeoff between recall and precision. Recall and accuracy are both enhanced by a large area under the curve; a low false positive rate is linked to a high resolution, and a high recall is linked to a low false negative rate. A more appropriate metric to use when assessing the efficacy of a model is its accuracy and recall curve.

Conclusion

Several datasets are incorporated into the methodology that we propose in this research for the purpose of finding DTIs. The model that has been presented is able to correctly predict drug target pairings by taking into account both the sequencing and the structural characteristics of those proteins. In contrast to the majority of the earlier techniques, which took into account evolutionary characteristics derived from proteins and amino acid sequences, the latter uses physical chemical properties and drugs. The field of drug discovery has seen the emergence of strong tools in the form of machine learning models, notably in the area of drug-target interactions (DTIs). The process of discovering new medicines has been revolutionised as a result of the capacity of these models to use large-scale biological and chemical data, in conjunction with modern algorithms. ML models are able to capture complicated interactions and give useful insights into the mechanism of action of medications. This is accomplished by the integration of a wide variety of data sources and the extraction of relevant information from pharmaceuticals and target proteins. These models speed up the process of drug development by automating the prediction process, which in turn reduces the reliance on experimental approaches that are both expensive and time-consuming. Even though there has been a lot of progress made in the creation of machine learning models for DTI prediction, there are still a few obstacles to overcome. The interpretation of model predictions, the confirmation of expected interactions through experimental research, and the need for larger and more diversified datasets are some of the challenges that need to be addressed. In order to effectively address these difficulties, it will be necessary for doctors, biologists, chemists, and computer scientists to work together across departments and disciplines.

References

[1] Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics 25, 2397–2403. doi: 10.1093/bioinformatics/btp433

[2] Cai, J., Cai, H., Chen, J., and Yang, X. (2018). Identifying “many-to-many” relationships between gene-expression data and drug-response data via sparse binary matching. IEEE/ACM Trans. Comput. Biol. Bioinform. 17, 165–176.

[3] Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. Bioinformatics 35, 602–610. doi: 10.1093/bioinformatics/bty662

[4] Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

[5] Cheng, L. (2019). Computational and biological methods for gene therapy. Curr. Gene Ther. 19, 210–210. doi: 10.2174/1566523219041910221133O7

[6] Cheng, L. (2020). Omics Data and Artificial Intelligence: New Challenges for Gene Therapy. Curr. Gene Ther. 20:1. doi: 10.2174/1566523220012006041500041

[7] Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. Brief. Bioinform. 22, 1442–1450. doi: 10.1093/bib/bbab042

[8] Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. Mol. Ther. Nucleic Acids 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

[9] Consortium, U. (2015). UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212.

[10] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046–1051.

[11] Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. Inform. Sci. 418, 546–560. doi: 10.1016/j.ins.2017.08.045

[12] Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

[13] Ding, Y., Tang, J., and Guo, F. (2020a). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. Knowl. Based Syst. 204:106254. doi: 10.1016/j.knosys.2020.106254

[14] Ding, Y., Tang, J., and Guo, F. (2020b). Identification of drug-target interactions via fuzzy bipartite local model. Neural Comput. Appli. 23, 10303–10319. doi: 10.1007/s00521-019-04569-z

[15] Fu, X., Cai, L., Zeng, X., and Zou, Q. J. B. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131