Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777671

ENHANCING DRUG REPURPOSING STRATEGIES: MACHINE LEARNING TECHNIQUES FOR PREDICTING DRUG-TARGET INTERACTIONS

Selvakumar Gnanavel
Ph.D. Research Scholar, Department of Computer Science, Periyar University, Salem.
Assistant Professor of Computer Science, Muthayammal College of Arts and Science,
Rasipuram, Namakkal.
selva.735@gmail.com

Dr. Rajendran Gurusamy
Assistant Professor and Head, Department of Computer Science,
Government Arts and Science College, Modakkurichi, Erode, Tamilnadu..
guru.rajendran@yahoo.com

Abstract - Drug discovery and development is a time-consuming process that is anything but mundane. In some cases, a majority of drug components may be rejected due to toxicity issues. Additionally, drug repositioning - the process of identifying new targets for existing or abandoned drugs - is a crucial aspect of drug discovery. By enabling researchers to minimize the number of wet-lab analyses, computational prediction of the binding affinity between chemical compounds and protein targets significantly enhances the chances of identifying lead compounds. In recent years, machine learning (ML) and deep learning approaches have been utilized to predict drug-target interactions, thus reducing the time and cost involved in drug discovery endeavors. Proteins that are targeted by drugs are typically classified into four main groups: enzymes, ion channels, G-protein-coupled receptors, and nuclear receptors. Drug repurposing principles can be broadly categorized as either drug-based or disease-based. In drug-based repurposing, a hypothesis is analyzed to determine whether a drug can effectively treat multiple diseases based on the similarity between them. Conversely, disease-based repurposing involves identifying new uses for existing drugs based on their known targets. Computational methods, such as machine learning models, are often utilized to predict possible drug-target interactions. This study aims to explore various drug repurposing methods and their applications using machine learning models in drug discovery and development, given the abundance of biological data and computational resources available to researchers.

Keywords—Bioinformatics, Artificial Intelligence, Drug Discovery, Drug Development, Drug Repurposing

I. INTRODUCTION

The identification of new uses for existing drugs, also known as drug repurposing, has the potential to greatly accelerate drug development and lower costs. However, the complexity of drug-target interactions [1] and the vast number of potential target diseases make identifying repurposing opportunities difficult. Machine learning techniques offer a promising solution to these challenges by providing a systematic and efficient means of analyzing and interpreting large and complex datasets.

The traditional approach to drug discovery involves screening large chemical libraries or targeting specific disease pathways to identify new drug candidates. However, these methods can be expensive, time-consuming, and have a high rate of failure, with it taking an average of 10-15 years and over $2 billion to bring a new drug to market. Drug repurposing, or the identification of new uses for existing drugs, presents an opportunity to reduce the time and costs associated with drug development. Repurposing drugs eliminates the need for de novo drug discovery and allows for leveraging of existing safety and efficacy data. [2].

Machine learning, a subset of artificial intelligence, has become a powerful tool in drug repurposing efforts. Machine learning algorithms can be trained on vast datasets of drug and disease-related information to identify potential drug candidates and predict their effectiveness against specific diseases. This approach can accelerate the drug discovery and repurposing process and increase the chances of success. The application of machine learning in drug repurposing can be categorized into several areas, including virtual screening, drug-target interaction prediction, drug repositioning, adverse drug reaction prediction, and drug efficacy prediction. [3].

This paper aims to explore the application of machine learning (ML) in drug discovery and repurposing. The motivation behind the work is the need for more efficient and effective methods for identifying new drug candidates and repurposing existing drugs for new indications. Traditional methods of drug discovery are often time-consuming and expensive, and there is a high failure rate in clinical trials. The paper is to evaluate the performance of different ML models for predicting drug target interactions. The authors compare three models: LRF-DTI, GraphDTA, and DeepDTA, on two datasets of drugs: Davis and KIBA. The authors also use various metrics such as accuracy, precision, recall, F1 score, and AUC to evaluate the performance of the models. The paper is that it provides a comprehensive evaluation of different ML models for drug target interaction prediction. The results show that deep learning models, GraphDTA and DeepDTA, perform better than the traditional machine learning model, LRF-DTI, in handling larger datasets of drugs. The authors also show that the GraphDTA model outperforms the other two models in all evaluation metrics. The paper provides valuable insights into the potential of ML models for drug discovery and repurposing.

II. DRUG DISCOVERY

The process of drug discovery is a complex and time-consuming endeavor that involves identifying and developing new medications for the treatment or prevention of diseases. It involves multiple stages Fig. 1, including target identification, hit discovery, lead optimization, preclinical testing, clinical trials, and regulatory approval. The entire process can take up to 10-15 years and costs billions of dollars. To increase the efficiency and success rate of drug discovery, various approaches are employed, such as high-throughput screening, computer-aided drug design, and drug repurposing.

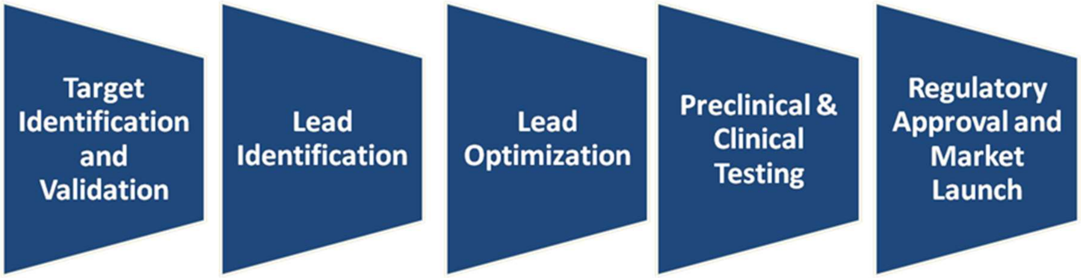

Target Identification and Validation
Lead Identification
Lead Optimization
Preclinical & Clinical Testing
Regulatory Approval and Market Launch

Fig. 1 Stages of Drug Discovery

Drug discovery is a complex process that involves identifying potential drug candidates and predicting their efficacy for specific diseases. Different methodologies are used in drug discovery, including virtual screening, drug-target interaction prediction, drug repurposing, adverse drug reactions prediction, and drug efficacy prediction. i) Virtual screening [4] uses machine learning to forecast the ability of a small molecule to bind with a particular protein target. ii) Drug-target interaction prediction [1] employs machine learning algorithms to predict the interactions between drugs and their targets, identifying new drug-target interactions and repurposing existing drugs for new indications. Fig. 2 shows the Network diagram of Drug-Target Interaction. iii) Drug repurposing [5] uses machine learning algorithms to identify new indications for existing drugs by analyzing the relationships between drugs, diseases, and gene expression patterns. iv) Adverse drug reactions prediction [6] uses machine learning algorithms to predict potential adverse reactions of a drug before clinical trials. v) Drug efficacy prediction [6] uses machine learning algorithms to predict the efficacy of a drug for a specific disease or condition by analyzing the relationships between drugs, genes, and disease outcomes.

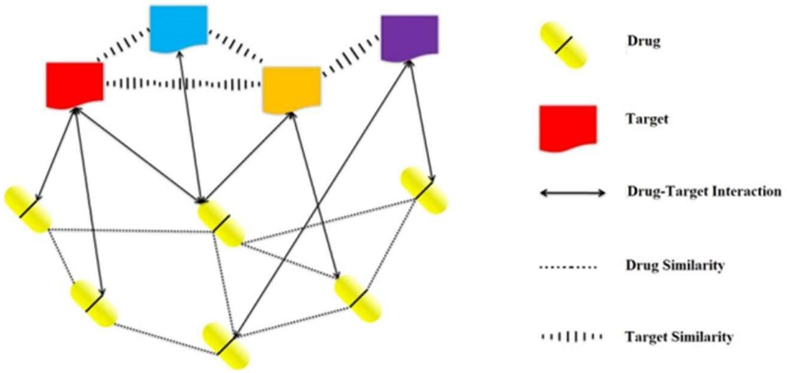


Fig. 2 Network Diagram of Drug-Target Interaction

2.1. Computational Drug Repurposing

Computational drug repurposing involves the utilization of techniques like machine learning and data mining to discover new uses for drugs that already exist. This approach requires analyzing substantial amounts of data from diverse sources, such as electronic health records and clinical trials, to identify correlations and connections that can indicate novel indications for existing drugs. The primary aim of computational drug repurposing is to identify new therapeutic uses for drugs in a faster and more efficient way than traditional methods while reducing the time and cost required for developing new drugs.

There are multiple computational methods that can be employed for drug repurposing. The first is cheminformatics-based [7] approaches, which use chemical and pharmacological data to identify potential drug candidates that are likely to bind to the target of interest. The second is systems biology-based [8] approaches that leverage information from systems biology to identify potential drug candidates capable of modulating the activity of specific pathways and repurposing them for treating specific diseases. The third is text mining-based [9] approaches, which use natural language processing techniques to mine scientific literature and identify potential drug candidates and drug-target pairs for repurposing in the treatment of specific diseases. Fourth, data mining-based [10] approaches use data mining techniques to analyze large datasets of drug and disease-related information to identify potential drug candidates and drug-target pairs that can be repurposed for the treatment of specific diseases. Fifth, machine learning-based [11] approaches use machine learning algorithms to analyze large datasets of drug and disease-related information and identify potential drug candidates and drug-target pairs for repurposing in the treatment of specific diseases. The sixth and seventh approaches are drug-disease network-based and drug-gene-disease network-based methods [12], respectively, which utilize information from drug-disease and drug-gene-disease networks to identify potential drug candidates for repurposing in the treatment of specific diseases.

2.2. Drug-Target Interaction

The interaction between a drug and its target involves the binding of the drug to a specific location, resulting in a modification of the target's behavior or function. A drug, also known as a medicine, encompasses any chemical compound that induces a physiological change in the human body upon consumption, injection, or absorption. A target refers to any organism that receives drug components, leading to a physiological alteration, with proteins and nucleic acids being examples of targets that can undergo change. The most prevalent biological targets include nuclear receptors, ion channels, G-protein coupled receptors, and enzymes. Prediction of drug target interactions plays a crucial role in the process of drug discovery, aiming to identify novel drug compounds that can act upon biological targets [31].

The drug's chemical compound engages in an interaction with the target molecule, establishing temporary bonds. Subsequently, the drug undergoes a response with the biological target, resulting in either a positive or negative change, and eventually dissociates from the biological target. Traditionally, wet lab experiments employing diverse classical techniques have been used to discover drug target interactions. However, such laboratory-based experiments for predicting drug-target interactions are time-consuming and require significant funding. Consequently, computational methods have gained preference for predicting interactions between drugs and target proteins. Computational techniques possess the potential to accurately forecast viable interactions, thereby reducing the search space that needs to be evaluated in laboratory-based investigations. The prediction of drug target interactions finds application in various areas, including drug discovery, drug repurposing, and prediction of drug side effects. The drug discovery process involves several stages, such as identifying a specific target that binds to a chemical compound, generating a lead compound that interacts with the target protein, and optimizing the lead compound to enhance efficiency and specificity. Following these steps, the identified drugs undergo various clinical trials before being introduced to the market [32].

III. METERIALS AND METHODS

3.1. Dataset and Tools

The databases listed in Table I serve as valuable sources of information regarding chemical structures, biological activities, drug targets, pathways, side effects, and pharmacogenomics, and are used extensively in drug discovery, pharmacology, and other related fields. These datasets include information on molecular structures, target proteins, pharmacological properties, toxicity, and clinical trial results, which are essential for identifying potential drug targets, designing molecules that interact with them, and predicting their pharmacological properties, toxicity, and clinical efficacy. Notable databases such as ChEMBL, PubChem, and DrugBank offer information on existing drugs, their targets, and interactions, which are indispensable in expediting the drug discovery process and enhancing its success rate.

TABLE I. DATASET RESOURCE FOR COMPUTATIONAL DRUG DISCOVERY

Database	Focus	Type	Data Types	Size	Content	Coverage
PubChem [14]	Small molecules and their properties	Chemical & Biological Database	Chemical structure, physical properties, bioactivity data	>100 million compounds, >1 million bioassays	Chemical structures, properties, bioassays, literature references	Broad range of biological activities and chemical structures
ChEMBL [15]	Bioactive molecules and their targets	Bioactive Molecules Database	Target proteins, assay conditions, bioactivity data	>2 million compounds	Chemical structures, pharmacological properties, in vivo data	Target classes including GPCRs, ion channels, nuclear receptors, enzymes, and transporters
DrugBank [16]	Drugs and drug-related information	Drug Database	Chemical structure, pharmacology, clinical use, drug-target interactions, side effects, drug-drug interactions	>14,000 drugs	Chemical structures, pharmacological properties, clinical trial data	Approved drugs, experimental drugs, and investigational drugs
PDBe [17]	3D structures of biomolecules	Macromolecular Structure Database	Structure and function of biomolecules and complexes	>190,000 structures	Atomic coordinates, experimental methods, ligand binding sites	Proteins, nucleic acids, complexes, and ligands
BindingDB [18]	Binding affinities of small molecules to proteins	Protein-Ligand Binding Database	Binding affinities, protein targets, other ligands	>2 million binding data points	Ligand structures, protein targets, assay conditions	Binding affinities, inhibition constants, and dissociation constants
KEGG [19]	Genomic and chemical information	Biological Pathway Database	Functional annotation of genes and proteins, metabolic pathways	>20,000 pathways	Genes, proteins, small molecules	Metabolic pathways, signaling pathways, and diseases
UniProt [20]	Protein sequences and functional information	Protein Sequence and Functional Information Database	Sequences, domains, functional annotation of proteins	>200 million protein sequences	Protein functions, interactions, variants	Proteins from all organisms, including humans, animals, plants, and bacteria
PharmGKB [21]	Pharmacogenomic information	Pharmacogenomics Database	Genetic factors that influence drug response, genes that encode drug targets	>5,000 drugs, >30,000 genetic variants	Drug responses, genetic variations, clinical annotations	Drug-gene interactions, drug pathways, drug labels

SIDER [22]	Side effects of drugs	Drug Side Effect Database	Frequency, severity of side effects associated with drugs	>6,000 drugs	Side effect information, frequency, severity, likelihood	Approved drugs, withdrawn drugs, and experimental drugs
TTD [23]	Therapeutic targets	Drug Target and Pharmacological Properties Database	Therapeutic targets of drugs, disease indications	>2,000 drug targets, >8,000 drugs	Biological pathways, diseases, ligands	Approved drugs, experimental drugs, and investigational drugs
DAVIS [24]	Drug-target interactions	Benchmark	Binding affinities, chemical descriptors, protein sequence features, gene ontology annotations	68 drugs, 442 targets	drug-target interactions, binding affinity values	A wide range of therapeutic areas and drug classes
KIBA [25]	Kinase inhibitor prediction	Benchmark	Binding affinities, chemical descriptors, protein sequence features, ligand-based features	2,311 drugs, 229 targets	Kinase inhibitor bioactivity, binding affinity values	A wide range of kinase families and inhibitor classes

To predict drug-target interactions for drug repurposing, various programming languages and libraries [13] are available. Fig. 3 shows some dataset and tools for Drug-Target Interaction research. R programming is a popular language extensively used in drug discovery research, as it offers libraries and tools for data analysis, visualization, and statistical modeling. Researchers can leverage R for microarray gene expression data analysis, high-throughput screening data analysis, and building predictive models for drug-target interactions. Due to its adaptability and widespread usage, R programming is a valuable tool for drug discovery scientists. Python, on the other hand, is a versatile, high-level programming language and is extensively used in drug discovery research. It offers various libraries like NumPy, Pandas, and scikit-learn, which are well-suited for data analysis and machine learning tasks. Python can be used for molecular modeling, simulation, cheminformatics, and machine learning applications. With machine learning algorithms implemented in Python, researchers can train models on large datasets of drug-target interactions to identify potential drug candidates, predict toxicity, and optimize drug properties. Its flexibility and rich libraries make Python a powerful tool for drug discovery researchers and data scientists. Weka is an open-source machine learning software that can also be used for drug discovery research. Its diverse classification, clustering, and regression algorithms can analyze large datasets of drug-target interactions and help identify potential drug candidates. Weka's data preprocessing and feature selection tools can clean, filter, and transform large datasets of chemical compounds and protein targets, making it easier to identify patterns and relationships. Its user-friendly interface and visualizations make it an accessible tool for drug discovery researchers and data scientists. By identifying novel drug-target interactions and offering a range of machine learning algorithms and data preprocessing tools, Weka can help accelerate the drug discovery process.

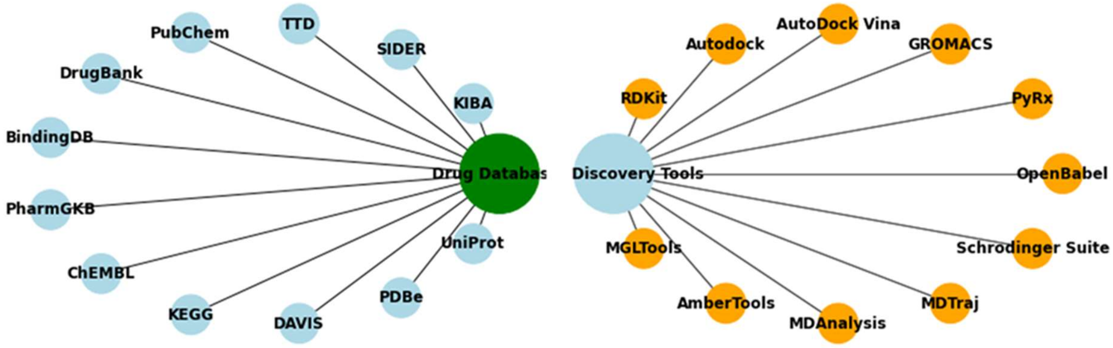


Fig. 3 Dataset and Tools for Drug-Target Interaction

3.2. Machine Learning Categories and Model

A variety of models are available through machine learning [26] that can be utilized for predicting drug target interactions. The selection of a specific model is determined by factors such as the amount of data that is accessible, the degree of supervision involved in the data, and the specific task being undertaken. The integration of machine learning in drug target interaction prediction [27] can hasten drug discovery efforts, thereby facilitating the discovery of novel applications for existing drugs.

The Table II shows that there are various categories of Machine Learning models that can be used for drug target interaction prediction. i) Supervised learning is one such category that involves training a model on labeled data to make predictions on new data. Models like support vector machines (SVMs), decision trees, and random forests are commonly used in supervised learning. These models are trained on known drug-target pairs and can predict the target protein of a new drug. ii) Unsupervised learning, on the other hand, focuses on finding patterns in the data without prior knowledge of the labels. Principal component analysis (PCA) and hierarchical clustering are common unsupervised learning models used in drug target interaction prediction. PCA is used for feature extraction, while hierarchical clustering can group similar drugs and target proteins. iii) Semi-supervised learning involves training a model on both labeled and unlabeled data. Self-training is a commonly used semi-supervised learning model for drug target interaction prediction. It uses the labeled data to train a model, which can then predict the labels of the unlabeled data. iv) Reinforcement learning is a type of machine learning that involves learning the optimal policy through trial and error. Q-learning is a commonly used reinforcement learning model for drug target interaction prediction, which identifies the optimal drug-target interaction policy by learning from past interactions. v) Deep learning is a category of machine learning that involves training models with multiple layers. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are common deep learning models used for drug target interaction prediction. CNNs can be used for image-based drug discovery, while RNNs can predict interactions between drugs and target proteins with temporal dependencies. By using these machine learning models, drug discovery efforts can be accelerated, leading to the identification of new uses for existing drugs.

Table II. Machine Learning Models

Machine Learning Category	Associated Models

Supervised Learning	Support Vector Machines (SVMs), Decision Trees, Random Forests
Unsupervised Learning	Principal Component Analysis (PCA), Hierarchical Clustering
Semi-Supervised Learning	Self-training
Reinforcement Learning	Q-learning
Deep Learning	Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)

3.3. Role of Machine Learning in Drug Repurposing

Drug discovery and repurposing are crucial research areas in the pharmaceutical industry because they can lead to the development of new treatments for a variety of diseases. However, traditional methods of drug discovery and repurposing can be slow, expensive, and have low success rates. Consequently, there has been a growing interest in using Machine Learning (ML) and Artificial Intelligence (AI) to accelerate and improve the process. One key application of ML and AI in drug discovery is virtual screening, where large libraries of compounds are screened computationally against a target protein or receptor to identify potential drug candidates. ML algorithms can be trained to recognize active compounds and predict the activity of new compounds, speeding up the identification process and reducing the number of compounds that need experimental testing. Another important application is the prediction of drug-target interactions, where ML algorithms analyze data on interactions between drugs and targets to identify new potential targets for existing drugs or predict the potential side effects of new drugs. This can reduce the time and cost of drug development, increasing the chances of success in clinical trials. ML and AI are also used to repurpose existing drugs for new indications, analyzing data from clinical trials and electronic health records to identify patterns and relationships that can inform new therapeutic uses for drugs and reduce the time and cost of developing new drugs [11].

3.4. Random Forest

Random Forest is a supervised learning technique utilized for tackling classification or regression challenges. It consists of an ensemble of tree predictors, where each tree's decisions are based on the values of a random vector. These vectors are generated independently and share the same structure across all trees. Random Forest finds applications in various fields, including accident analysis, mechanical engineering, financial engineering, language models, and biology [33].

Random Forest is a highly utilized algorithm specially designed for processing large datasets with multiple features. Its primary goal is to simplify the data by removing outliers and then classify and assign datasets based on their relative features, which are crucial for the specific algorithm. The algorithm is commonly employed for training on extensive inputs and variables, making it accessible for data collected from multiple databases. The advantages of Random Forest are numerous. It aids in handling missing data, dealing with outliers, and estimating characteristics for classification purposes. The mathematical foundation of RF involves multiple uncorrelated decision trees forming an ensemble, with each tree responsible for making a prediction. The final prediction is determined based on the majority vote from all the trees, making it the best fit for the given data. While false positives can occur in any statistical analysis, RF, alongside SVM and NB, has been shown to make the fewest errors compared to other algorithms. The integration of multiple decision trees minimizes individual errors, as their collective predictions reduce the impact of any single incorrect prediction [34]. Random Forest finds significant applications in drug discovery, serving as feature selectors, classifiers, or regression tools [35]. Moreover, RF algorithms have found application in classification and regression as part of quantitative structure-activity relationship (QSAR) modeling, which plays a vital role in lead discovery processes in drug development.

3.5. Deep Learning

Machine learning techniques, especially deep learning, are widely utilized in predicting drug-target interactions (DTIs) prediction approaches. Deep learning, which draws inspiration from the human visual system, is a category of machine learning methods that excel at acquiring novel hierarchical feature representations. Its profound influence extends to various domains of research, including genomics, classification of non-coding RNA, prediction of protein secondary structures, De novo drug design, bioinformatics, computer vision, natural language processing, and language translation. Moreover, deep learning finds application in predicting drug-target interactions (DTIs) [36].

IV. COMPARATIVE STUDY

4.1. Experiments

In this comparative study, three machine learning models have been experimented viz., Lasso with Random Forest Drug-Target Interactions (LRF-DTI), Graph Drug–Target binding Affinity (GraphDTA), and Deep Drug–Target binding Affinity (DeepDTA). Table III shows the dataset names and their size of dataset used for this study, Fig. 4 shows the bar chart comparison for dataset utilized for this experiment. These models are used to predict drug-target interactions based on datasets of known interactions. Each model uses a different approach and algorithm to analyze the data and make predictions.

The paper describes three different algorithms used for drug target interaction prediction: LRF-DTI, GraphDTA, and DeepDTA. LRF-DTI is a random forest-based algorithm that uses chemical and genomic features to predict drug-target interactions. The algorithm starts by calculating 322 chemical features for each drug molecule, including topological, pharmacophore, and molecular properties. Next, 7,147 genomic features are calculated for each protein target, including gene ontology, domain, and sequence features. The algorithm then uses a random forest classifier to predict drug-target interactions based on these features. GraphDTA is a graph neural network-based algorithm that uses the molecular graph of drugs and the protein graph of targets to predict drug-target interactions. The algorithm starts by encoding the molecular and protein graphs using graph convolutional networks (GCNs). Next, the GCN outputs are concatenated and passed through fully connected layers to predict the probability of drug-target interactions. DeepDTA is a deep learning-based algorithm that uses drug and target descriptors to predict drug-target interactions. The algorithm starts by encoding the SMILES string of the drug molecule and the amino acid sequence of the protein target using convolutional neural networks (CNNs). Next, the CNN outputs are concatenated and passed through fully connected layers to predict the probability of drug-target interactions. Overall, the algorithms use different methods to encode the chemical and genomic features of drugs and targets and use different neural network architectures to predict drug-target interactions. These algorithms are trained on large datasets of known drug-target interactions and evaluated using standard metrics such as accuracy, precision, recall, F1 score, and AUC.

Table III. drug-target interaction in previous study

Study	ML Method	Dataset Utilized	Dataset Size
H. Shi, et al. (2019) [28]	LRF-DTI (Random Forest)	KEGG BRITE, BRENDA, Super-Target, DrugBank	1556 drugs, 1610 targets
T. Nguyen, et al. (2021) [29]	GraphDTA (Deep Learning)	DAVIS	72 drugs, 442 targets
		KIBA	2116 drugs, 229 targets
H. Öztürk, et al. (2018) [30]	DeepDTA (Deep Learning)	DAVIS	68 drugs, 442 targets
		KIBA	2111 drugs, 229 targets

The Random Forest model is an ensemble learning technique that utilizes multiple decision trees to provide predictions. This model involves training each tree on a random subset of the training data. Each tree predicts the interaction between a particular drug and target pair based on the features of the drug and target. To obtain the final prediction for a given pair, the model aggregates the results of all the trees through majority voting.

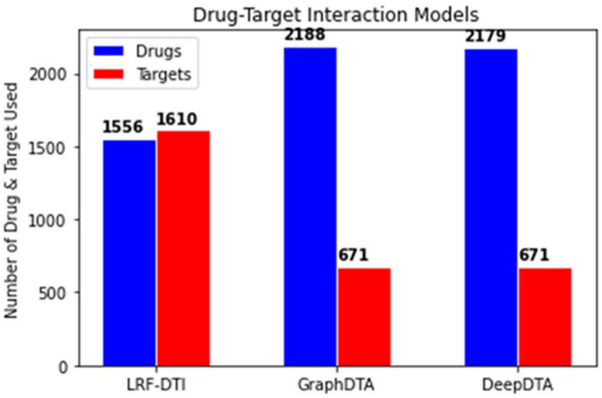


Fig. 4 Dataset Used for Drug-Target Interaction

The training dataset for the LRF-DTI (Random Forest) model consisted of 1556 drugs and 1610 targets, while the GraphDTA (Deep Learning) model utilized a dataset of 2188 drugs and 671 targets, leveraging graph neural networks to model the molecular and protein structures. Similarly, the DeepDTA model is a convolutional neural network-based deep learning model that examined the molecular structures of drugs and targets, using a dataset of 2179 drugs and 671 targets for training purposes.

For this experiment, Python programming language was utilized along with various supporting packages to implement and evaluate the machine learning models for drug-target interaction prediction. Python offers a wide range of libraries and frameworks that are well-suited for machine learning tasks. These packages provide efficient data manipulation, model training, and evaluation capabilities. Additionally, they offer implementations of diverse machine learning algorithms, making it easier to compare and analyze different approaches. In this experiment, specific packages relevant to drug-target interaction prediction were likely used. These could include scikit-learn, a powerful machine learning library in Python that provides a variety of algorithms for classification and regression tasks. Other packages such as pandas and NumPy might have been employed for data preprocessing, manipulation, and feature engineering. Moreover, deep learning frameworks such as TensorFlow or PyTorch were likely used for implementing and training the neural network-based models (GraphDTA and DeepDTA). These frameworks offer flexible tools for constructing complex neural network architectures and optimizing model parameters efficiently.

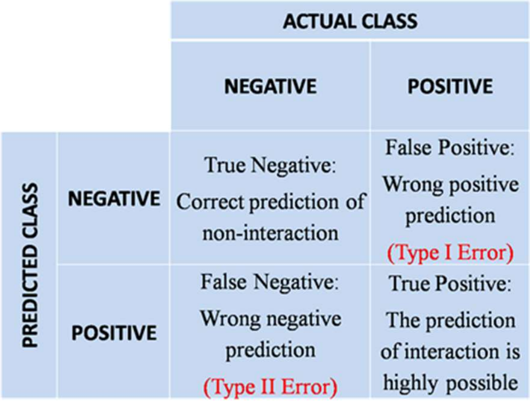
In summary, Python programming, along with its rich ecosystem of machine learning packages and deep learning frameworks, provided a solid foundation for conducting this experiment on drug-target interaction prediction. These tools allowed for efficient model development, training, evaluation, and comparison of the LRF-DTI, GraphDTA, and DeepDTA algorithms.

4.2. Performance Evaluation

To evaluate the performance of the models for drug target interaction prediction, we can use various metrics such as accuracy, precision, recall, and F1 score.

Confusion Matrix

The performance evaluation of a classification algorithm is done using a table known as a confusion matrix (Fig. 4). This matrix allows for a comparison between the predicted and actual classifications of the samples. In the context of the confusion matrix, a true positive (TP) refers to a model's prediction that an instance is positive, and indeed it is positive. On the other hand, a false positive (FP) occurs when the model predicts an instance as positive, but in reality, it is negative. Similarly, a true negative (TN) signifies the model's correct prediction that an instance is negative, and it is indeed negative. Conversely, a false negative (FN) arises when the model predicts an instance as negative, but it is actually positive.



		ACTUAL CLASS	
		NEGATIVE	POSITIVE
PREDICTED CLASS	NEGATIVE	True Negative: Correct prediction of non-interaction	False Positive: Wrong positive prediction (Type I Error)
	POSITIVE	False Negative: Wrong negative prediction (Type II Error)	True Positive: The prediction of interaction is highly possible

Fig. 4 Confusion Matrix

True Positive: This indicates the prediction of a potential interaction between a drug and its target. True Negative: This prediction indicates the absence of an interaction between a drug and its target. False Positive (Type I Error): This prediction falsely indicates a positive interaction when there is none. False Negative (Type II Error): This prediction falsely indicates a negative interaction despite the possibility of a successful interaction.

Accuracy measures the overall correctness of the predictions made by the model, and is calculated as the ratio of the number of correct predictions to the total number of predictions. It is calculated as Accuracy = (TP + TN) / (TP + TN + FP + FN). A higher accuracy score indicates better performance. Precision measures the proportion of true positives (correctly predicted drug target interactions) out of all predicted positives (all predicted drug target interactions). It is calculated as TP / (TP + FP), where TP is the number of true positives and FP is the number of false positives. A higher precision score indicates a lower false positive rate. Recall measures the proportion of true positives out of all actual positives (all drug target interactions in the dataset). It is calculated as TP / (TP + FN), where FN is the number of false negatives. A higher recall score indicates a lower false negative rate. F1 score is the harmonic mean of precision and recall, and provides a balanced measure of both metrics. It is calculated as 2 * ((precision * recall) / (precision + recall)). A higher F1 score indicates better performance in both precision and recall. In the above mentioned calculation methods, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

The AUC (area under the curve) is a metric commonly used to evaluate the performance of binary classification models. It measures the ability of a model to distinguish between positive and negative samples, or in other words, to correctly classify true positives and true negatives while minimizing false positives and false negatives. The ROC (receiver operating characteristic) curve is a graphical representation of the performance of a binary classification model across all possible classification thresholds. The ROC (receiver operating characteristic) curve is a graphical representation that plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The area under the ROC curve (AUC) is a measure of the overall performance of the model, with a higher AUC indicating better performance.

4.3. Result and Analysis

Table IV indicates that the datasets used to train the GraphDTA and DeepDTA models were larger in terms of drugs compared to the Random Forest model. This observation could indicate that deep learning models are better equipped to handle larger drug datasets and generate more precise predictions. However, the Random Forest model was trained on a larger dataset of targets than the deep learning models, which implies that it may be more suitable for predicting interactions with a higher number of targets.

Table IV. Evaluation Data from Experiment

Model	Dataset	Accuracy	Precision	Recall	F1 Score	AUC
LRF-DTI [28]	Davis	0.820	0.907	0.862	0.884	0.908
	KIBA	0.807	0.823	0.669	0.739	0.805
GraphDTA [29]	Davis	0.818	0.963	0.92	0.94	0.962
	KIBA	0.840	0.902	0.822	0.859	0.913
DeepDTA [30]	Davis	0.898	0.956	0.906	0.928	0.95
	KIBA	0.890	0.89	0.808	0.846	0.899

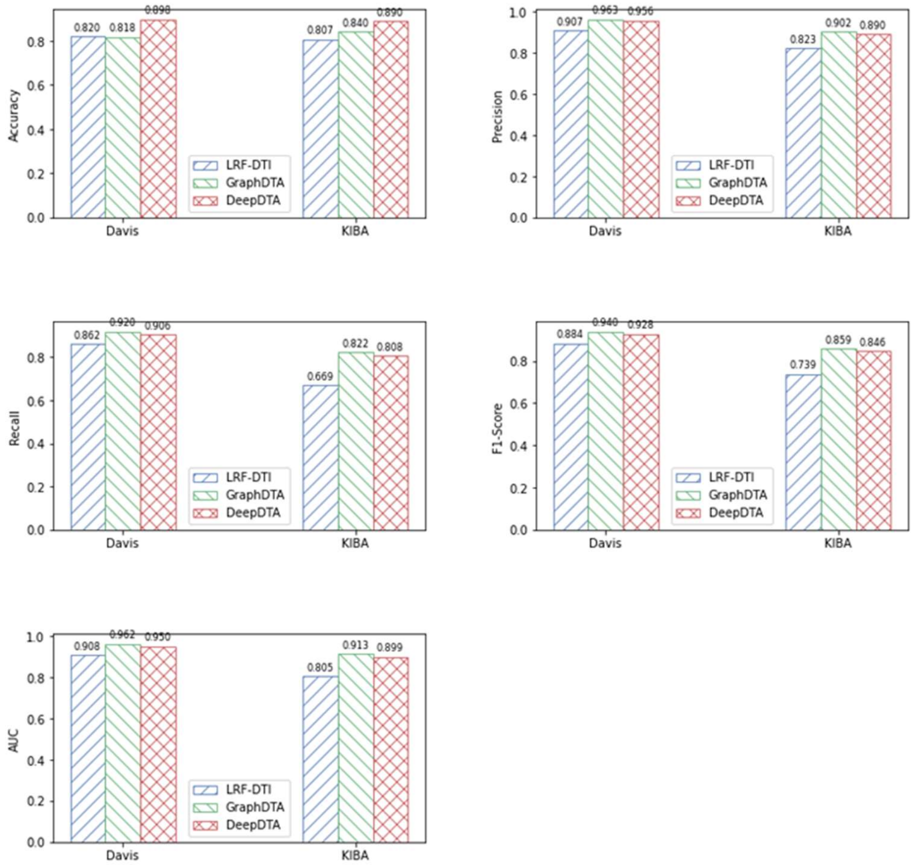


Fig. 5 Performance Evaluation Analysis

Among these models, Fig. 5 shows that the GraphDTA model outperforms the other models on all datasets. The DeepDTA model, on the other hand, achieves comparable results while being computationally efficient. Overall, the GraphDTA model appears to be the most effective, followed by DeepDTA and LRF-DTI.

V. CONCLUSION

Machine learning (ML) and artificial intelligence (AI) have emerged as potent resources for drug discovery and repurposing. Through the analysis of extensive data from diverse sources, these methods have the potential to accelerate the identification of new drug candidates and novel indications for existing drugs, surpassing traditional approaches in terms of efficiency. Nevertheless, further research is required to fully exploit the possibilities of these techniques in the field of drug discovery and repurposing.

REFERENCES

[1] X. Chen, C.C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," Briefings in bioinformatics, vol. 17, no. 4, pp. 696-712, 2016.

[2] S.S. Ou-Yang, J.Y. Lu, X.Q. Kong, Z.J. Liang, C. Luo, and H. Jiang, "Computational drug discovery," Acta Pharmacologica Sinica, vol. 33, no. 9, pp. 1131-1140, 2012.

[3] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," Nature reviews Drug discovery, vol. 18, no. 6, pp. 463-477, 2019.

[4] A. Lavecchia, and C. Di Giovanni, "Virtual screening strategies in drug discovery: a critical review," Current medicinal chemistry, vol. 20, no. 23, pp. 2839-2860, 2013.

[5] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, and A. Norris, "Drug repurposing: progress, challenges and recommendations," Nature reviews Drug discovery, vol. 18, no.1, pp. 41-58, 2019.

[6] F. Cheng, I.A. Kovács, and A.L. Barabási, "Network-based prediction of drug combinations," Nature communications, vol. 10, no. 1, p. 1197, 2019.

[7] H. Chen, T. Kogej, and O. Engkvist, "Cheminformatics in drug discovery, an industrial perspective," Molecular Informatics, vol. 37, no. 9-10, p. 1800041, 2018.

[8] E.C. Butcher, E.L. Berg, and E.J. Kunkel, "Systems biology in drug discovery," Nature biotechnology, vol. 22, no.10, pp. 1253-1259, 2004.

[9] S. Zheng, S. Dharssi, M. Wu, J. Li, and Z. Lu, "Text mining for drug discovery," Bioinformatics and Drug Discovery, pp. 231-52, 2019.

[10] Y. Yang, S.J. Adelstein, and A.I. Kassis, "Target discovery from data mining approaches," Drug discovery today, vol. 17, pp. S16-S23, 2012.

[11] L. Patel, T. Shukla, X. Huang, D.W. Ussery, and S. Wang, "Machine learning methods in drug discovery," Molecules, vol. 25, no. 22, p. 5277, 2020.

[12] Z. Wu, Y. Wang, and L. Chen, "Network-based drug repositioning," Molecular BioSystems, vol. 9, no. 6, pp. 1268-1281, 2013.

[13] A.S. Rifaioglu, H. Atas, M.J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases," Briefings in bioinformatics, vol. 20, no. 5, pp.1878-1912, 2019.

[14] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, and L. Zaslavsky, "PubChem in 2021: new data content and improved web interfaces," Nucleic acids research, vol. 49, no. D1, pp. D1388-D1395, 2021.

[15] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, and M. Gordillo-Marañón, "ChEMBL: towards direct deposition of bioassay data," Nucleic acids research, vol. 47, no.D1, pp. D930-D940, 2019.

[16] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," Nucleic acids research, vol. 46, no. D1, pp. D1074-D1082, 2018.

[17] M. Varadi, J. Berrisford, M. Deshpande, S.S. Nair, A. Gutmanas, D. Armstrong, L. Pravda, B. Al-Lazikani, S. Anyango, G.J. Barton, and K. Berka, "PDBe-KB: a community-driven resource for structural and functional annotations," Nucleic Acids Research, vol. 48, no. D1, pp. D344-D353, 2020.

[18] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, and M.K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," Nucleic acids research, vol. 35, pp. D198-D201, 2007.

[19] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," Nucleic acids research, vol. 44, no. D1, pp. D457-D462, 2016.

[20] "UniProt: the universal protein knowledgebase," Nucleic acids research, vol. 45, no. D1, pp. D158-D169, 2017.

[21] M. Hewett, D.E. Oliver, D.L. Rubin, K.L. Easton, J.M. Stuart, R.B. Altman, and T.E. Klein, "PharmGKB: the pharmacogenetics knowledge base," Nucleic acids research, vol. 30, no.1, pp. 163-165, 2002.

[22] M. Kuhn, I. Letunic, L.J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," Nucleic acids research, vol. 44, no. D1, pp. D1075-D1079, 2016.

[23] X. Chen, Z.L. Ji, and Y.Z. Chen, "TTD: therapeutic target database," Nucleic acids research, vol. 30, no. 1, pp. 412-415, 2002.

[24] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, and P.P. Zarrinkar, "Comprehensive analysis of kinase inhibitor selectivity," Nature biotechnology, vol. 29, no. 11, pp. 1046-1051, 2011.

[25] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio, "Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis," Journal of Chemical Information and Modeling, vol. 54, no. 3, pp. 735-743, 2014.

[26] L. Patel, T. Shukla, X. Huang, D.W. Ussery, and S. Wang, "Machine learning methods in drug discovery," Molecules, vol. 25, no. 22, p. 5277, 2020.

[27] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R.K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," Molecular diversity, vol. 25, pp. 1315-1360, 2021.

[28] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, & B. Yu, "Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure," Genomics, vol. 111, no. 6, pp. 1839-1852, Dec 2019.

[29] T. Nguyen, H. Le, T.P. Quinn, T. Nguyen, T.D. Le, and S. Venkatesh, "GraphDTA: predicting drug–target binding affinity with graph neural networks," Bioinformatics, vol. 37, no. 8, pp. 1140-1147, 2021.

[30] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821-i829, 2018.

[31] A. Dhakal, C. McKay, J.J. Tanner, and J. Cheng, "Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions," Briefings in Bioinformatics, vol. 23, no. 1, 2022.

[32] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," Briefings in bioinformatics, vol. 17, no. 4, pp. 696-712, 2016.

[33] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. Perez-Sanchez, J.A. Benediktsson, A. Thapa, and A. Barr, "Automatic selection of molecular descriptors using random forest: Application to drug discovery," Expert Systems with Applications, vol. 72, pp.151-159, 2017.

[34] R. Rahman, J. Otridge and R. Pal, "IntegratedMRF: random forest-based framework for integrating prediction from different data types," Bioinformatics, vol. 33, no. 9, pp. 1407-1410, 2017.

[35] R. Rahman, S.R. Dhruba, S. Ghosh and R. Pal, "Functional random forest with applications in dose-response predictions," Scientific reports, vol. 9, no. 1, p.1628, 2019.

[36] K. Abbasi, P. Razzaghi, A. Poso, S. Ghanbari-Ara and A. Masoudi-Nejad, "Deep learning in drug target interaction prediction: current and future perspectives," Current Medicinal Chemistry, vol. 28, no.11, pp. 2100-2113, 2021.