Drug Target Interaction Prediction Using Machine Learning Techniques – A Review

A. Suruliandi[1], T. Idhaya[1], S. P. Raja[2] *

[1] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli, TamilNadu (India)
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, TamilNadu (India)

Received 13 August 2021 | Accepted 4 January 2022 | Early Access 10 November 2022

Abstract

Drug discovery is a key process, given the rising and ubiquitous demand for medication to stay in good shape right through the course of one's life. Drugs are small molecules that inhibit or activate the function of a protein, offering patients a host of therapeutic benefits. Drug design is the inventive process of finding new medication, based on targets or proteins. Identifying new drugs is a process that involves time and money. This is where computer-aided drug design helps cut time and costs. Drug design needs drug targets that are a protein and a drug compound, with which the interaction between a drug and a target is established. Interaction, in this context, refers to the process of discovering protein binding sites, which are protein pockets that bind with drugs. Pockets are regions on a protein macromolecule that bind to drug molecules. Researchers have been at work trying to determine new Drug Target Interactions (DTI) that predict whether or not a given drug molecule will bind to a target. Machine learning (ML) techniques help establish the interaction between drugs and their targets, using computer-aided drug design. This paper aims to explore ML techniques better for DTI prediction and boost future research. Qualitative and quantitative analyses of ML techniques show that several have been applied to predict DTIs, employing a range of classifiers. Though DTI prediction improves with negative drug target pairs (DTP), the lack of true negative DTPs has led to the use a particular dataset of drugs and targets. Using dynamic DTPs improves DTI prediction. Little attention has so far been paid to developing a new classifier for DTI classification, and there is, unquestionably, a need for better ones.

Keywords

Chemogenomics, Drug Databases, Drug Discovery, Drug Target Interactions, Machine Learning, Targets, Target Databases.

DOI: 10.9781/ijimai.2022.11.002

I. Introduction

Discovering new drugs is critical and driven by the need for medication in daily life, partly brought on by changing environmental conditions. Nevertheless, drug discovery is not easy, it demands time as well as money, and the drug success rate is usually low. Computer-Aided Drug Design (CADD) is considered a computational discipline that aims to discover, design, and develop therapeutic chemical targets. There are 3 phases in drug design - discovery, development, and registry.

In the first phase, discovery, the focus is on identifying a new drug and its targets, based on binding sites. The second phase, development, involves pre-clinical research, where the drug is tested on animals for safety. Successful research means that human trials are set in motion. In the third phase, registry, the Food and Drug Administration (FDA) thoroughly reviews all the submitted drug-related data and decides on its approval or otherwise. Initiating an efficient computational model that finds potential Drug Target Interaction (DTI) from biological data helps understand the biological process, recognize novel drugs, and offer improved therapeutic medicine for illnesses of all sorts. Drug development has three trial phases, each of which is more expensive than the others. As of today, the cost of drug development has risen from US$3.4 million to US$8.6 million and US$21.4 million for phase I, phase II and phase III trials, respectively [1]. A new drug could fail to pass the test in any of the three drug development trial phases, notwithstanding the expense, effort and time involved.

* Corresponding author.
E-mail addresses: suruliandi@yahoo.com (A. Suruliandi), idhayathomas003@gmail.com (T. Idhaya), avemariaraja@gmail.com (S. P. Raja).

II. State of the art Methods

DTI is the process of finding new drugs and targets for drug development. Drug and target molecules are discovered through their interactions. Drug discovery methods are ligand-based, docking-based and chemogenomics-based, and involve parameters like biomarker identification, structure unavailability, physique and condition, and environmental factors. Current research is focused on maximizing interactions so the drugs formulated can successfully treat disease. The new drugs developed today, though based on knowledge of existing ones, could still have adverse side effects. Incidentally, a drug developed for a particular disease may be used, quite unexpectedly, to treat another disease with no side effects whatsoever, a process

Please cite this article in press as:
A. Suruliandi, T. Idhaya, S. P. Raja. Drug Target Interaction Prediction Using Machine Learning Techniques – A Review, International Journal of Interactive Multimedia and Artificial Intelligence, (2022), http://dx.doi.org/10.9781/ijimai.2022.11.002

referred to as drug repurposing [2], [3]. It is essential in drug discovery to establish the interaction between a drug and a target gene. The docking-based method needs a 3D structure of the target protein or gene for the process to work. The success of a newly developed drug depends on how well it fares in the market, particularly in terms of whether the purpose for which it was originally designed is being fulfilled. The possibility of successfully identifying DTI is enhanced by working on binding factors or interacting sites. This is a difficult process, given the limited information on drugs and targets. Bioinformaticians have tried to draw information from factors driving drugs and targets. The automated tools employed to improve the success rate by discovering more interactions or binding sites between drugs and their targets are intended to actively assist doctors and bioinformaticians. Scientists today work in drug development using ML predictive analysis techniques to understand drugs and targets, thus boosting DTI success prediction.

A. Drug Developing Procedure

Drugs are synthesized chemicals that control, prevent, and cure and diagnose illnesses. Disease diagnosis is carried out through reading the body's reactions to drug molecules in the form of positive biological responses. In pharmacological terms, the biomolecule whose function and activity are modified by a specific drug is termed the drug target. Biomolecules can be proteins, nucleic acids, receptors, enzymes, and ion channels. The DTI process interacts or binds the drug molecule to the active biomolecule site with the same structural or functional properties as the drug molecule, culminating in the creation of a new product as in Fig.1. The human body assimilates the product, resulting in a cure.

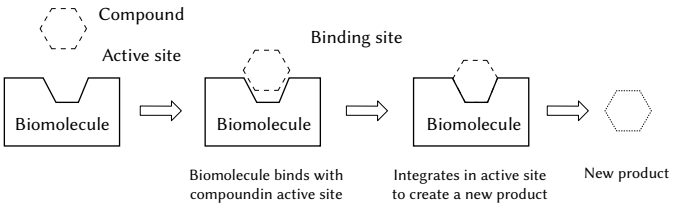

Fig.1. Drug Developing Procedure.

Drugs are developed in three phases. In the first phase, a drug and its target are discovered by means of the interacting or binding site, using substrate on the active site of protein. In the second phase, the drug is subjected to animal testing for safety's sake. In the third phase, the drug has human trials, following which it is marketed.

B. In-Silico Approaches in Drug Discovery

In-vitro is a technique where the process of drug discovery takes place in a controlled environment but not within a living organism. Here a pool of potential compounds is identified and narrowed down to find most reliable compound for treatment. In-vivo is a technique where the process of drug discovery takes place within a living organism by giving the reliable compounds to the human trials. Both the data collected from in-vitro and in-vivo are given as input features to the in-silico methods for drug prediction, which is a computational method. The computational DTI prediction method is categorized into the three approaches [4].

1. Docking-Based Approach

A docking-based approach in DTI prediction requires a 3D structure for simulation. Consequently, it is not applicable where a large number of proteins is involved, as in, for instance, the G-Couple Protein Receptor and ion channel, whose structures are far too complex to be obtained. The simulation is significant in regard to the time taken and its overall efficiency.

2. Ligand-Based Approach

A ligand-based approach works on the premise that a drug can be predicted without the 3-Dimensional structure of targets and with the existing knowledge of drugs and its targets.

3. Chemogenomics-Based Approach

A chemogenomics-based approach integrates both the chemical space of drugs and the genomic space of targets into a single pharmacological space. The challenge here is that there are too few DTI pairs and too many unknown interaction pairs.

C. Motivation and Justification

The in-vitro prediction of DTI from biological data calls for a lot of effort in the search for new drugs and targets. Identifying potential drugs and targets is a painstaking step in initiating drug discovery. Despite the plethora of research on DTI prediction in the recent past, prediction is still material-intensive and protracted. Predicting interaction between DTPs continues to challenge researchers. The motivation for this review is to help researchers in the drug development domain access state-of-the-art methods used in ML for DTI predictions, and so enhance the quality of research. To this end, several insightful articles on DTI procedures and methods that help discover new drugs and targets differently are reviewed. The machine learning (ML) techniques used to predict DTIs are studied, each with its strengths and limitations. The research is categorized, based on the ML techniques used in the prediction. Thereafter, it is qualitatively and quantitatively analyzed to understand ML and DTI better so the latter can be improved.

The contributions of this paper are as follows, Articles related to ML and DTI in drug development are studied in detail and categorized, based on the machine learning techniques deployed as in section III. The feature selection techniques used in DTI prediction suggest the best features for use. Articles on DTI prediction using ML techniques have described how ML manages datasets from miscellaneous databases, balances imbalanced data, handles large-scale datasets and features and, finally, examines at length the ML algorithms used in DTI prediction. Articles that are qualitatively analysed in section V based on ML techniques to understand their strengths and weaknesses. A quantitative analysis in section VI follows to find the most appropriate classifiers for DTI predictions.

D. Organization of the Paper

The paper is organized as follows. Section II provides an overview of state of the art methods involved in DTI prediction using ML techniques. In Section III, Machine learning techniques used for DTI prediction are summarized. In Section IV, databases used for DTI prediction are discussed. In Section V, a qualitative analysis of the ML techniques used for DTI is presented. In Section VI, a quantitative analysis of DTI prediction methods is offered. Section VII discusses DTI prediction. Section VIII concludes the study and offers new directions for future research.

III. Machine Learning (ML) Techniques Used for DTI Prediction

Computational models use ML techniques for prediction because they optimize data better and perform better as well. ML techniques, which learn data without relying on previously defined formulas, are grouped into two – supervised and unsupervised learning. Supervised learning predictions are based on observed existing knowledge from known data, while unsupervised learning predictions do the same without. Predictions are guesses based on existing knowledge from the data at hand. On the other hand, Classification refers to the process

of differentiating between known and unknown labels. The objective of this paper is to explore ML techniques involved in improving DTP identification to find DTIs. The identification of a new drug involves the drug and its target. Because of large number of features of both drugs and targets manually extracting them would be a time taking process, so the researchers use only tools like ChemCPP, EDragon, CDK, Open Babel, RDkit, PADEL for extracting the features from drugs and Protr, SPICE, Propy, ProtDcal, ProtParam for extracting features from targets. Drug and target features are extracted and concatenated with each other to form DTPs. The pairs are analyzed for interaction prediction; specifically, to observe whether or not the DTPs interact. The ML techniques analyzed are explained qualitatively and quantitatively and the classifier used for DTI prediction is found. The DTI prediction here mainly uses a static database. Prediction can be improved when there are more targets and drugs with the interaction between them yet to be ascertained. In recent times, CADD has been used to develop drugs for immunodeficiency syndrome, influenza virus infection, glaucoma and lung cancer [5]. CADD helps in pharmacological, Pharmacodynamics and in-silico toxicity prediction, which identifies or filters inactive or toxic molecules [6] and naturally gets ML involved in DTI prediction strategies [7]-[10]. Thus to improve drug development various methods based on drugs and targets are developed using ML techniques. Fig.2 shows DTI prediction through ML techniques with targets and drugs taken from diverse databases. Drug and target features are extracted using a slew of tools or web servers. Subsequently, the most influential features alone are selected and used for DTI prediction with several ML classifiers to complete the process.

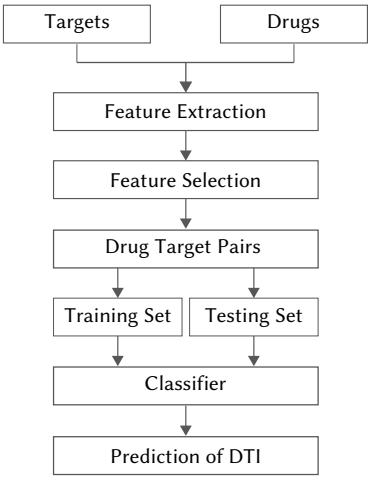

Fig.2. Flow of DTI Prediction.

In-silico methods include Machine learning, Data mining, Network analysis tool and data analysis tool, Quantitative Structure Analysis Relationship (QSAR), pharmacophores, homology modeling, Here Machine learning technique is more feasible than all other methods for working with drug discovery data for analysis. The trending research in drug discovery is "Identification of screening hits (compounds)" which helps in finding the particular compounds target with more potency at different level like binding, reducing the side effects, efficiency, and also increases the life of patients by changing the function of the biomolecule.

A. Chemogenomics-Based Machine Learning (ML) Techniques for DTI Prediction

The chemogenomics-based prediction approach is computationally predicted using ML-based, graph-based or network-based methods. ML-based methods are explained below in Fig 3.

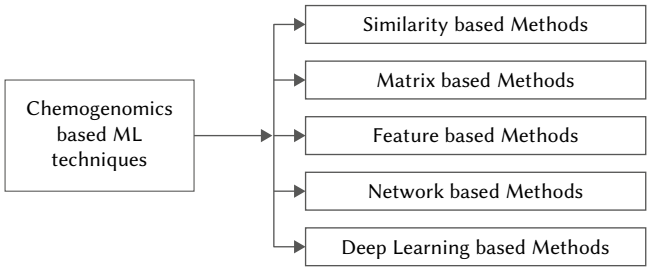

Fig.3. Chemogenomics based ML Techniques.

1. Similarity-Based Methods

The most commonly used DTI prediction methods use drug and target similarity measures in tandem with the distance between each pair of drugs and its targets [11]-[18]. These methods use the drug, target and drug-target interaction similarity scores based on prior knowledge of their interaction similarity. The similarity is obtained using a distance function like the Euclidean. For instance, if the following function is employed for the nearest neighbor algorithm, assuming two vectors x1 and x2, the distance between the vectors is found using equation (1) as D(x1, x2) where

$$D\ (x1, x2) = 1 - \frac{x1.x2}{\|x1\|\|x2\|} \tag{1}$$

and the same dimension and distance are calculated using the Euclidean norm and the inner product. The similarity between a drug and a target is given through the pharmacological similarity of the drug, the genomic similarity of the protein sequence, and the topological properties of a multipartite network of previously known drug-target interaction knowledge. The disadvantage of these methods is that they use knowledge drawn from a small quantum of labelled data, while there exist large quanta of unlabeled data.

2. Matrix-Based Methods

Several studies [19]-[24] have shown that matrix-based methods outperform the rest in DTI prediction. The interaction matrix is

$$X_{mxn} = \begin{bmatrix} x11 & \cdots & x1n \\ \vdots & \vdots & \vdots \\ xm1 & \cdots & xmn \end{bmatrix} \tag{2}$$

For i=1: m and j=1: n,

$$X_{ij} = \begin{cases} 1 \text{ if drug i and target j interact} \\ 0 \text{ there is no interaction between drug i and target j} \end{cases}$$

The first move in DTI prediction is to break down matrix X_{mxn} into two matrices, Y_{mxk} and Z_{nxk}, where $X \sim YZ^T$ with $k < m, n$, and where Z^T denotes the swapped matrix of Z. This process of factorizing matrices in lower order makes it easier for matrix-based approach to deal with the missing data. With these methods, however, the distance between the drug and target appears to be the same and establishes the strength of the interaction between them, embedding them in a low-dimensional matrix. The reliability of these methods is affected when the drug and target data increase in volume, impacting the capacity to find their interaction.

3. Feature-Based Methods

Feature-based prediction methods largely use the support vector machine to find drug-target interaction [25]-[33]. Any pair of targets and drugs may be represented with features, leading to binary classification or two-class clustering with positive or negative interactions. Features are represented as F

$$F = \{d + t\},\ d = d_1, d_2, d_3, \ldots d_a \text{ and } t = t_1, t_2, t_3, \ldots t_b \tag{3}$$

where d denotes the drug features of length a and t the target features of length b, respectively.

4. Network-Based Methods

Network-based methods [34]-[40], which use graph-based techniques to predict DTI, are considered simple and reliable interaction prediction methods. Here, the drug-drug similarity, target-target similarity and known interactions between DTI are integrated into a heterogeneous network, operating on the simple logical principle that similar drugs interact with similar targets.

5. Deep Learning-Based Methods

Deep learning-based approaches can reduce the loss of feature information in predicting DTIs. However, they need adequate information to predict interaction and drug repurposing [41]-[45]. The two steps of deep learning include generating feature vectors and predicting interaction. The target property and drug property generate a features matrix for prediction.

IV. Databases Used in DTI Prediction

Interaction prediction demands the twin data items of drugs and targets, and a working knowledge of their interaction. The popular databases used in this study fall into two categories, drug-centered and target-centered. More than 20 databases associated with interaction prediction are not directly involved in DTI prediction, though the data contained therein maybe used as input for prediction. The popular database, KEGG, used here for prediction, is divided into the sub-databases of KEGG BRITE [46] and KEGG DRUG [47], incorporating a mass of biological data from genes and proteins.

A. Chemical European Molecular Biology Laboratory (ChEMBLdb)

The data gathered is a chemical database of bioactive molecules [48] which are collected from numerous literature studies. With millions of chemical compounds, 10,000 drugs and 12000 targets, the ChEMBLdb was established by the EMBL – European Bioinformatics Institute in 2002.

B. Chemical – Protein Annotation Resource (ChemProt)

ChemProt [49] has Chemical-Protein interactions data that integrates data from multiple databases of chemical protein annotations. It comprises data from the PDSP, DrugBank, PharmGKB, PubChem and STITCH databases. ChemProt also integrates therapeutic effects, adverse drug reactions and chemical-biological disease data.

C. Drug Gene Interaction Database (DGIdb)

This database has information on Druggable targets with their effects and drug-gene interaction data [50].

D. DrugBank

DrugBank is one of the most well-known databases in DTI study, with details about drug-like compounds, their different forms, target genes and side effects brought on by drug intake. The DTI data in this database that have been collected from an array of literature studies has extensive commercial uses [51].

E. Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG is an out-of-the-box database with exhaustive details of genes and genome sequences [52]. The KEGG databases are divided into four categories. The first has three numbers of databases KEGG - BRITE, PATHWAY and MODULE. The second has four databases that carry genomic information– KEGG-GENOME, KEGG-GENE, KEGG-SSDB and KEGG ORTHOLOGY. The third has five databases with chemical information KEGG- COMPOUNDS, KEGG-REACTION, KEGG-RCLASS, KEGG-ENZYME and KEGG-GLYCAN. The fourth has four databases carrying health information– KEGG-DISEASE, KEGG-DRUG. The comprehensive KEGG has a wealth of DTI information and outclasses others.

F. Library of Integrated Network-Based Cellular Signatures (LINCS)

This database holds information on the KINOME scan. Kinases are small molecule-binding assays that help study the interaction between drug compounds for testing purposes. The database consists of 398 datasets on fluorescence imaging, ELISA and ATAC-sequence data [53].

G. PROMISCUOUS

The database has network-based drug repositioning data with information on drugs, proteins and the side effects of every drug. The information on protein is from the Unitprot database, while details on drugs and side effects are from the SuperDrug and Sider databases, incorporated into the *LINCS* [54].

H. Search Tool for Interacting Chemicals (STITCH)

STITCH has information on target or protein interaction with small molecules, collected from PubChem databases and literature studies [55].

I. SuperTarget

SuperTarget is a web resource that carries information on DTIs, drug metabolic rate, pathways, and Gene Ontology (GO) terms, as well as on adverse medical side effects. The DTI information is sourced from PubMed, DrugBank, KEGG, PDB and TTD, and potential drug-target relationships are extracted from Medline [56].

J. Therapeutic Target Database (TTD)

The Target Therapeutic Database has therapeutic information on protein and nucleic acid, assimilated from literature studies and miscellaneous databases with DTI data [57].

K. BRENDA -The Comprehensive Enzyme Information System (BRENDA)

This is an enzyme database with information on enzyme-ligand interaction. The data collected is drawn from literature studies based on enzyme nomenclature [58].

L. Drug Central

Drug Central is a Food and Drug Association (FDA)-approved drug database. The database incorporates relevant information on drugs in the form of structure, bioactivity and regulatory records, which are categorized as small molecule active ingredients and biological active ingredients [59].

M. Protein Drug Interaction Database (PDID)

Protein Drug Interaction Database (PDID) has DTI for all the structural proteome for human beings, with predictions made using the ILbind, SMAP and eFindSite software[60].

N. Pharos

Pharos is the user interface for giving knowledge about Illuminating Druggable Genome (IDG) to the knowledge management center for three of the protein families like GPCR, Ion Channel and Kinases [61].

O. PubChem

PubChem [62] has information about chemical substances and their biological activity. The PubChem database incorporates three databases–Substances, Compounds and BioAssay. The first stores data on chemical information, the second has exclusive chemical structures obtained from substances, while the third holds biological information on the extracted substances.

P. Super Drug

Super Drug [63] offers information on all drug features collected from several databases and incorporated here. The database has 2-Dimensional and 3-Dimensional structure information on small molecule drugs, side effects and drugs pharmacokinetics specifications.

Q. FDA Adverse Event Reporting System (FAERS)

The FDA Adverse Event Reporting System (FAERS) is a database with information obtained from adverse events and medication error reports submitted to the FDA on side effects, as well as keywords for drugs [64].

R. SIDe Effect Resource (SIDER)

SIDER is a database [65] that holds data on marketed medicines and their side effect information, including frequency of side effects, and also drug and its side effect classification.

S. International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) -The IUPHAR/BPS Guide to Pharmacology

The IUPHAR/BPS is considered as a guide to pharmacology [66] is an open access knowledge website that provides information on licensed drugs and their targets and holds information on small molecule drugs.

T. Cancer Drug Resistance Database (CancerDR)

CancerDR offers elaborate information on anti-cancer drugs and their pharmacological profiling. CancerDR helps in effective personalized cancer therapies and identifies gene-encoding drug targets, based on genetic and residual resistance [67].

U. Binding Database (DB)

Binding DB is a binding database that holds the DTI of small molecules as well as all the interaction data collected from an array of literature studies. This is an extensive database for protein ligand binding affinity [68].

V. ZINC is not Commercial (ZINC)

ZINC is the largest database [69] comprising every drug needed for new ligand discovery. Information on drugs and the targets they can interact with are collected here. ZINC is a major database for researchers looking for the chemical composition of their biological targets.

W. Psychoactive Drug Screening Program (PDSP)

The Psychoactive Drug Screening Program (PDSP) [70] screens compounds with previous reports of pharmacological, biochemical and behavioural activity. It is chiefly used to identify novel targets in the treatment of mental disorders.

X. A Summary of Databases

Table I, summarizes the general statistical information on every database.

TABLE I. Databases Involved in Dti Prediction

S. No	Databases	No. of Targets	No. of Drugs	No. of Interactions
1	ChEMBL	12482	1879206	15504603
2	ChemProt	20000	170000	-
3	DGI db	41100	9495	29783
4	DrugBank	5175	13338	26932
5	KEGG	19711	4948	260000
6	LINCS	1469	41847	-
7	PROMISCUOUS	6548	5258	23702
8	STITCH	9600000	430000	-
9	SuperTarget	6000	196000	330000
10	TTD	3101	34019	-
11	BRENDA	84000	20500	-
12	Drug Central	-	4543	-
13	PDID	3746	5100	-
14	Pharos	20244	130166	-
15	PubChem	79622	96157016	-
16	Super Drug	4456	4605	-
17	FAERS	-	24842	-
18	SIDER	1430	140064	-
19	IUPHAR/BPS	1396	1105	443
20	Cancer Dr	-	148	-
21	Binding DB	7020	489416	1132739
22	Zinc	-	20 million	-
23	PDSP	738	7449	-

V. Qualitative Analysis of Machine Learning Techniques for DTI Prediction

Qualitative analysis helps in an understanding of the ML techniques involved in DTI predictions, based on the quality and characteristics of the methods used. Qualitative analysis outcomes are descriptive, and inferences are drawn easily from the data obtained and the analysis of DTI prediction is shown in Table III-VII.

TABLE II. Dataset Used in Dti Prediction

Dataset	Targets	Drugs	DTI
Enzyme	664	445	2926
Ion Channel	204	210	1476
GPCR	95	223	635
Nuclear Receptor	26	54	90

GPCR- G-Protein Coupled Receptor

The Yaminishi et al. [71] Bench Mark (BM) dataset has been the only one used by many of the researchers for the purpose because it incorporates diverse drug and target data to create a new DTI dataset. The BM dataset is shown in Table II.

A. Review of Literature for Similarity-Based Methods

Similarity-based methods consider similarities between drugs and targets to identify DTIs. Perlman et al. [11] proposed a scheme that incorporates multiple drugs and targets similarity to predict DTI using the logistic regression SITAR (Similarity-based Inference of drug-TARgets) framework. Mei et al. [12] proposed a bipartite local model (BLM)-based method to handle the candidate problem of baseline BLM-NII (BLM with Neighbor-based Interaction profile Inferring). Van Laarhoven and Marichiori [13] developed a weighted nearest neighbor (WNN)algorithm that directly uses the GIP (Gaussian interaction profile) kernel by drawing up a profile of the interaction score for a new drug (WNN-GIP). Shi et al. [14] proposed a method to handle missing interactions using a cluster of similar targets that is Super Target Clustering (STC). Buza K [15] proposed a K-nearest neighbor (KNN)-based method with hubness-aware classification and error correction to maximize the detrimental effect of bad hubs (EcKNN-KNN with error correction). Zhang et al. [16] posited a framework that develops a drug-drug linear neighbourhood, calculates the similarities, and predicts drug-target interaction profile and label propagation (LPLNI-Label Propagation with Linear Neighbourhood Information). Zhang et al. [17] developed a clustering algorithm by incorporating

TABLE III. Qualitative Analysis of the Articles Using Similarity-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference 11 (2011)	Logistic Regression (LR)	250 Proteins, 315 Drugs	-	Wrapper Feature Selection	10 Fold CV	Lists the selected features	Only 10 features are considered	Targets of 307 drugs are predicted
Reference 12 (2012)	Bipartite Local Model-NII (BLM-NII)	BM Dataset	-	-	LOOCV and 10 Fold CV	NII procedure for finding drugs and targets	Whenever new drug or target is given as input it is not considered as there is no training data	57 % of DTI has been predicted
Reference 13 (2013)	Weighted Nearest Neighbor (WNN)	BM Dataset	-	-	LOOCV and 5 Fold CV	Uses regularized least square algorithm to find the new drug based on the old drugs	No difference between indirect and direct targets. These are not measured to interact with drugs.	Prediction of DTI interaction which show top 5 prediction for each dataset.
Reference 14 (2015)	Super Target Clustering (STC)	BM Dataset	-	-	5 Fold CV	Finds missing interaction using cluster of targets.	Considers only about missing interaction not more about existing DTI	Finds new drugs and targets and potential interaction
Reference 15 (2016)	K-Nearest Neighbor (KNN)	BM Dataset	Finger print extraction for drugs	-	5 Fold CV, LOOCV	Hubness awareness and ensemble size gives high accuracy	LOOCV over fits and then shifted to 5 Fold CV	Improved prediction of DTIs around 12 prediction is found
Reference 16 (2017)	LPLNI	BM Dataset	-	-	LOOCV	Integrating similarities of different features	Considers only fingerprint as features for drugs	A promising tool for DTI prediction
Reference 17 (2017)	Multi-View DTI	1253drugs, 887 targets	-	-	20 trials of 5 Fold CV	Enrichment analyzes of drugs and targets	No details of experiments	56 newly identified clusters
Reference 18 (2018)	K-Nearest Neighbor (KNN)	BM Dataset	-	-	5 trials of 10 Fold CV	Calculating probability based weight and similarity based weight for targets	Considers only Ranking of top several integrations of drug and targets	34 % better prediction than previous methods

BM Dataset - Bench Mark dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation.

TABLE IV. Qualitative Analysis of The Articles Using Matrix-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference 19 (2009)	BRDTI	BM Dataset	-	-	5 Trials of 10 Fold CV	Incorporates target bias and context alignment for drug and target similarities	More survey based on DTI is to be done for better prediction.	DTI leads to Drug repurposing and adverse drug reaction prediction
Reference 20 (2012)	KBMF	BM Dataset	-	-	5 Fold CV	Interaction score is generated using factorization methods	Better for only 12 low dimensional projection	Similarity based DTIs.
Reference 21 (2017)	MLRE	608 protein, 326 drugs, 114 interactions	Structural view and chemical view of drug are extracted	-	5 Fold CV	Preserving the point wise linear regression	Noisy observation leads to disagreement data	Predict interaction based on chemical view with SVM and graph based methods
Reference 22 (2017)	VB-MK-LMF	BM Dataset	-	-	5 Trials of 10 Fold CV	DTI matrices are linked to weighted common observations	Works well for mid-sized datasets	DTI predicted
Reference 23 (2018)	Pseudo SMR	BM Dataset	Extraction of Pseudo AAC	-	5 Fold CV	Uses extremely randomized tree methods and it is computationally more efficient	Uses only Pseudo AAC Descriptors.	Predicted 15 Potential DTIs.

BM Dataset - Bench Mark Dataset, CV- Cross Validation, AAC –Amino Acid Composition.

TABLE V. Qualitative Analysis of the Articles Using Feature-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference 25 (2011)	Regularized Least Square	BM Dataset	-	-	LOO CV and 5 Trials of 10 Fold CV	Combining GIP with target kernel and drug kernel	Increase kernel with more information about DTI	15 known interaction was predicted
Reference 26 (2016)	Krons-Regularized Least Square	BM Dataset	Replaced missing values with mean of data	-	5 Fold CV	Incorporates both known and unknown interaction and make a general purpose learner	Balancing the data is not considered	Prediction of interval as measure of confidence
Reference 27 (2016)	Weighted SVM	BM Dataset	Structural similarity, Gene Function similarity was extracted	-	5 Fold CV	Finds some unlabeled sample as negative sample and also considers positive samples beneath unlabeled samples	Asks for using structure but we cannot get structure for all the targets	Predicts Interaction and listed 3 top known interaction
Reference 28 (2016)	Ensemble learning	5877Drugs 3348Targets 12674DTI	PROFEAT for Target and Rcpi for Drug	-	5 Fold CV	Ensemble learning to address issues of class imbalance	Oversampling is done which increases noise	Predicted more than 20 Known DTI
Reference 29 (2017)	Discriminate Vector Machine	BM Dataset	AAC feature were Extracted	Principal Component Analysis (PCA)	5 Fold CV	Uses LBP histogram vectors which retains evolutionary information of amino acid	Only AAC information is used for prediction	Not listed the predicted DTI
Reference 30 (2017)	Support Vector Machine	BM Dataset	-	-	10 Fold CV	Multiple Kernel combination is used for prediction	GIP based prediction	Compound-Protein-Interaction
Reference 31 (2017)	REP Tree Algorithm	2719 E 1372 IC 630 GPCR 86 NR	-	-	10 Fold CV	Considers different families of proteins by using various learning rate	No cross validation is done	DTI prediction
Reference 32 (2017)	Adaboost	BM Dataset	PSSM for target and SMILE for drug were extracted	Sequential Forward Feature Selection (SFFS)	5 Fold CV	Balanced Data using RUS and CUS techniques	Not considered domain features	Listed top 10 known interaction
Reference 33 (2018)	Bagging based ensemble	5877Drugs 3348Targets 12674 DTI	PROFEAT for Target and Rcpi for Drug	-	10 Fold CV	Considered class imbalance and used Neighbourhood balanced bagging for balancing the data and active learning strategy is used	Not discussed about Features	14 out of 16 known interactions have been detected.

BM dataset - Bench Mark dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation, PROFEAT-PROtein FEATures, AAC- Amino Acid Composition, Rcpi-R package for extracting features for compound protein interaction..

drug and target data from structural and chemical viewpoints with existing knowledge of interactions (MDTI- Multiview DTI). Shi and Li [18] advanced an improved Bayesian ranking DTI method that adds weights for unknown drugs and targets using weighted neighboring drugs and targets (WBRDTI–Weighted Bayesian Ranking DTI).

B. Review of Literature for Matrix-Based Methods

Matrix-based methods use matrix similarity for DTI prediction. Rendle et al. [19] proposed an algorithm based on the Bayesian Personalized Ranking (BPR) matrix factorization which incorporates drug and target similarities to predict DTIs (BPRDTI). Gonen [20] proposed a method to factorize the matrices with interaction score matrix so as to find new drugs and targets and determine their interaction using kernelized Bayesian matrix factorization (KBMF). Li et al. [21] introduced an algorithm to find a low-rank representation embedding (LRE) technique and fix errors in point wise linear reconstruction. This was done to obtain a different view of the structural and chemical features of drugs and targets as Single view LRE and Multiview LRE, respectively (LRE). Bolgar et al. [22] developed a method integrating multiple kernels, weights, and graphs, all regularized to model the probability of DTI prediction (VB-MK-LMF). Huang et al. [23] propounded an extension of the structure activity relationship classification by implementing the extremely randomized tree (ERT) using the pseudo substitution matrix representation (SMR) of the target (Pseudo-SMR). Marta et al. [24] proposes a local model-agnostic for interaction prediction.

TABLE VI. Qualitative Analysis of the Articles Using Network-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference 34 (2012)	Network based Inference (NBI)	BM Dataset	-	-	10 Fold CV	Used a bipartite graph for prediction	Imbalanced data is used	5 new DTI were predicted
Reference 35 (2012)	Network-based Random Walk with Restart on the Heterogeneous network (NRWRH)	BM Dataset	-	-	LOOCV	Used RWR to get potential DTI using bipartite graph network	Leaves the target which has no drug it is considered ass zero matrix	29 new DTI were predicted
Reference 36 (2013)	Network-Consistency-based Prediction Method (Net CBP)	BM Dataset	-	-	Not discussed Properly	DTI predicted using bipartite graph network	Considered as zero matrix	Listed out several DTI
Reference 37 (2015)	Normalized Multi information Fusion	BM Dataset	-	-	Not discussed properly	Integrates robust PCA with biological information	In order to improve performance more negative dataset to be built to find the interactions.	Predicts interaction
Reference 38 (2015)	Random Walk Restart (RWR)	467Targets 544Drugs	-	-	-	RWR on heterogeneous network using chemical features	Considered only fingerprints features for drugs	110 drugs predicted for 3419 targets
Reference 39 (2018)	IN - Random Walk with Restart (RWR)	12015 Drug 1895445 Target	-	Principal Component Analysis (PCA)	5 Fold CV	Used both labelled and unlabeled data for prediction	Data is imbalanced	Predicts interaction between drug and targets
Reference 40 (2019)	Neighbourhood Regularized Logistic Matrix Factorization (NRLMF)	BM Dataset	Calculates similarities of drugs and targets	-	10 Fold CV	Improved using rescoring matrix	Not more parameters are considered	Predicts interaction but not listed

BM Dataset - Bench Mark Dataset, CV- Cross Validation, LOOCV-Leave One Out Cross Validation.

C. Review of Literature for Feature-Based Methods

Feature-based methods consider drug and target features for DTI prediction. Van Laarhoven et al. [25] proposed an algorithm that integrates the DTI network information with the Gaussian Interaction Profile kernel using the Regularized Least Square (RLS). Ezzat et al. [26] developed a framework for DTI prediction using the voting of the decision tree, random forest, STACK and Laplacian Eigen base classifiers, and also considered imbalanced classes for prediction. Nascimento et al. [27] advanced a method that incorporates both known and unknown interaction data using the RLS. Lan et al [28] developed a framework for DTI prediction by taking unlabeled samples using the weighted SVM (PUDT-Positively Unlabeled Drug Targets). Li et al. [29] proposed a method to find DTIs as a structure activity relationship (SAR) classification with the principal component analysis (PCA), using the Discriminative Vector Machine (DVM). Ohue et al. [30] proposed an approach that uses virtual screening and the Pairwise Kernel Method (PKM). Zhang et al. [31] proposed an ensemble-based approach for a random projection ensemble (RPE) of the REP tree algorithm (Drug RPE). Rayhan et al. [32] developed a model using targets in the form of a matrix (position-specific scoring matrix - PSSM) and drug molecules features for DTI prediction using the AdaBoost classifier (iDTI-EsBoost). Sharma and Rani [33] proposed an ensemble (Bagging-Ensemble) model that uses active learning methodology to predict DTIs (BE-DTI).

D. Review of Literature for Network-Based Methods

These methods use networks of similar drugs and targets for DTI prediction. Cheng et al. [34] proposed a bipartite Network Based Inference (NBI) method for DTI prediction. Chen et al. [35] developed an RWR framework to get potential DTIs using a bipartite graph network (NRWRH-Network-based Random Walk with Restart on Heterogenous network). Chen et al. [36] used this method for both labelled and unlabeled data DTI prediction (NETCBP-Network Consistency-based Prediction). Peng et al. [37] proposed a method that incorporates the PCA to reduce dimensions and integrate data from multiple drug and target sources for DTI prediction (NMIF-Normalized Multi-Information Fusion). Seal et al. [38] proposed a model that needs matrix inversion and score of relevance between two nodes in a weighted graph of DTIs (RWR-Random Walk with Restart). Huang et al. [39] proposed a 2-network-based rank algorithm that involves the random walk and bipartite graph (IN-RWR-intra network with Random Walk). Ban et al. [40] developed a method based on improving the NRLMF algorithm by calculating the NRLMF scores as the expected beta distribution values. Beta distribution value is calculated using the interaction information and NRLMF score (NRLMF-beta).

TABLE VII. Qualitative Analysis of the Articles Using Deep Learning-Based Methods

Source	ML Tech	Dataset	Pre processing/ Feature Extraction	Feature Selection	Validation	Strength	Weakness	Outcome
Reference 41 (2017)	Deep DTI	1520 Targets 1412 Drugs 12524 samples	-	-	10 Fold CV	Uses DBN and Fine tune RBM in greedy way.	Only known interaction are used	DTI probability which are useful for drug repurposing
Reference 42 (2018)	Deep DTA	442 Targets 68 Drugs 30056 DTI	-	-	Concordance index	Creating CNN blocks of targets, drugs	Predefined features are considered for CNN blocks of protein	Predicts binding affinity
Reference 43 (2018)	AUTO DNP	BM Dataset	PSSM for Target and PubChem fingerprint has taken for drugs	-	5 Fold CV	Uses Auto encoder blocks to create Deep NN	Only CTD descriptors are considered.	Predicts interaction
Reference 44 (2019)	LASSO – DNN	3546 Proteins 5834 Drugs 14792 DTI	-	-	10 Fold CV	Considers Tripeptide composition feature of proteins	More number of functions are used.	Diseases treated by drug and its association with breast cancer is listed
Reference 45 (2019)	Deep Convolution-DTI	3675Targets 11950Drugs 32,568 DTI	-	t-distributed stochastic neighbor embedding (t-SNE)	5 Fold CV	Similarity acts as a informative descriptors	Considers only CTD descriptors of targets	Predicts interaction

BM Dataset - Bench Mark Dataset, CV- Cross Validation, CTD – Composition, Transition and Distribution, PSSM - Position Specific Scoring Matrix, PubChem – PubChem is a Chemical Information database.

E. Review of Literature for Deep Learning-Based Methods

Deep learning-based methods use the drug and target features for DTI prediction. Wen et al. [41] proposed a method that takes raw target and drug features using a deep belief network (DBN) and predicts DTI in drugs approved by the Food and Drug Association (DeepDTIs). Ozturk et al. [42] proposed a DTI prediction model using target sequences and drug molecule to predict drug target binding affinity (DeepDTA). Wang et al [43] developed a computational model using a stacked auto encoder for DTI prediction (AUTO-DNP). You et al. [44] presented a method based on protein and drug features with LASSO regression model in tandem with the deep neural network (DNN) to predict DTI (LASSO-DNN). Lee et al. [45] proposed a DTI prediction model using local protein residue patterns in DTI (DeepConv-DTI).

VI. Quantitative Analysis of Machine Learning Techniques in DTI Prediction

Quantitative analysis is applied to determine the best prediction performance method, using different ML techniques with appropriate metrics. The prediction method must deal with the steps of data pre-processing and feature selection, as well as drug and target integration. The best machine learning prediction method includes the hyper parameters and association index for DTI prediction. Of the various ML techniques [11]-[44] available, the best is chosen for prediction. Tables X-XIV depict the quantitative analysis of the results of several ML methods in DTI prediction that help enhance performance.

A. Performance Metrics

A confusion matrix is used to calculate performance measures from test set values in terms of true positives, true negatives, false positives and false negatives among classes that are to be classified as integrates or not integrates. Table VIII shows the confusion matrix for DTI and Table IX the performance metrics used. Integrates here refers to drugs that produce a positive DTP result, that is, the integrating drug can be used to treat a target it integrates with. The converse is true with non integrates, which refers to drugs that produce a negative DTP result, that is, the non integrating drug cannot be used to treat a target it does not integrate with.

TABLE VIII. Confusion Matrix

	Integrates	Non Integrates
Integrates	True Positive	False Positive
Non integrates	False Negative	True Negative

TABLE IX. Performance Metrics Used in DTI Prediction

S. No	Metrics Used	Formula	Metrics Description
1.	Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	Accuracy is the ratio of correct prediction out of total number of predictions
2.	Sensitivity/ Recall	$TP/(TP+FN)$	Measure of quantity
3.	Precision	$TP/(TF+FP)$	Measure of quality
4.	AUC	False Positive vs. True Positive	Curve shows the relation between False Positive and True Positive
5.	AUPR	Precision vs. Recall	Curve shows the relationship between the Precision and Recall
6.	MCC	$\frac{(TP*TN)-(FP*FN)}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}}$	Mathew's Correlation Coefficient
7.	F1 Score	$TP/(TP+1/2+TP/(FP+FN))$	Harmonic average of Precision and Recall

TABLE X. Quantitative Analysis of the Similarity-Based Methods Used in DTI Prediction

Similarity Based Methods																					
S. No	ML Tech.	Accuracy				Sensitivity/ Recall				Precision/nDCG				AUC				AUPR/MAP			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	LR	-	-	-	-	-	-	-	-	-	-	-	-	92.2	92.7	94.6	86.3	87.7	88.9	93.9	85.1
2.	BLM-NII	-	-	-	-	-	-	-	-	-	-	-	-	98.8	99.0	98.4	98.1	92.9	95.0	86.5	86.6
3.	WNN	-	-	-	-	-	-	-	-	-	-	-	-	81.9	75.5	84.8	78.8	29.9	24.9	30.8	43.4
4.	STC	-	-	-	-	-	-	-	-	-	-	-	-	81.2	81.1	87.5	87.1	38.5	36.7	41.4	53.3
5.	KNN	-	-	-	-	-	-	-	-	-	-	-	-	95.4	97.2	97.2	-	83.7	85.5	62.8	-
6.	LPLNI	-	-	-	-	-	-	-	-	-	-	-	-	97.0	97.6	99.4	99.1	90.6	94.6	96.8	94.9
7.	Multi-view DTI	-	-	-	-	-	-	-	-	-	-	-	-	86.9				-	-	-	-
8.	KNN	-	-	-	-	-	-	-	-	nDCG				98.3	98.4	96.2	94.8	MAP			
										90.8	95.9	94.0	94.5					88.0	94.2	91.5	92.7

E-Enzyme, IC-Ion Channel, G-G-Protein Coupled Receptor (GPCR), N-Nuclear Receptor, AUC-Area Under Curve, AUPR-Area Under Precision Recall, nDCG-normalized Discounted Cumulative Gain PPV-Positive Predicted Values, MCC-Mathew's Correlation Coefficient, MAP-Mean Average Precision.

TABLE X. Quantitative Analysis of the Matrix-Based Methods Used in DTI Prediction

Matrix based Methods																					
S. No	ML Tech.	Accuracy				Sensitivity/Recall				Precision/nDCG				AUC				AUPR/MCC			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	BRDTI	-	-	-	-	-	-	-	-	nDCG				98.1	98.2	95.5	92.3	-	-	-	-
										89.7	95.3	92.9	94.8								
2.	KBMF	-	-	-	-	-	-	-	-	-	-	-	-	83.2	79.9	85.7	82.4	-	-	-	-
3.	MVLRE	-	-	-	-	-	-	-	-	-	-	-	-	65.0	51.4	61.7	-	-	-	-	-
4.	VB-MK LMF	-	-	-	-	-	-	-	-	-	-	-	-	98.7	98.9	97.6	95.7	89.0	91.0	80.0	77.0
5.	Pseudo SMR	89.4	87.8	82.9	83.3	89.5	87.9	82.1	95.2	90.2	87.8	82.1	76.3	96.0	93.8	90.5	96.3	MCC			
																		81.8	78.7	71.8	71.6

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient, nDCG-normalized Discounted Cumulative Gain.

TABLE XII. Quantitative Analysis of the Feature-Based Methods Used in DTI Prediction

Feature Based Methods																					
S. No	ML Tech.	Accuracy/PPV/MCC/ F1 Score				Sensitivity/Recall				Precision				AUC				AUPR/MCC/ F1 Score			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	RLS	-	-	-	-	-	-	-	-	-	-	-	-	98.2	98.5	94.5	88.7	88.1	91.8	70.0	60.4
2.	Krons-RLS	-	-	-	-	-	-	-	-	-	-	-	-	97.9	98.7	95.1	92.4	-	-	-	-
3.	Weighted SVM	PPV				24.0	14.0	16.0	7.0	99.0	99.0	94.0	97.0	88.4	83.1	87.8	88.5	-	-	-	-
		36.0	74.0	58.0	64.0																
4.	Ensemble Learning	-	-	-	-	-	-	-	-	-	-	-	-	90.0				-	-	-	-
5.	DVM	93.1	91.7	89.3	92.2	92.9	92.6	89.2	96.6	93.1	90.9	89.4	88.6	92.8	91.7	88.56	93.00	MCC			
																		86.3	83.4	78.77	84.80
6.	REP Tree	94.0	91.0	88.0	88.0	92.0	89.0	81.0	87.0	90.0	86.0	83.0	79.0	98.0	97.0	94.0	93.0	F1 Score			
																		91.0	88.0	82.0	83.0
7.	Adaboost	MCC				85.0	84.0	84.0	87.0	85.0	78.0	80.0	92.0	96.0	93.0	93.0	92.0	68.0	48.0	50.0	79.0
		18.0	29.0	26.0	22.0																
		F1 Score																			
		10.0	20.0	19.0	24.0																
8.	BE-DTI	-				88.0				-				92.7				88.6			

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

TABLE XIII. Quantitative Analysis of the Network-Based Methods Used in DTI Prediction

Network Based Methods																					
S. No	ML Tech.	Accuracy				Sensitivity/Recall				Precision				AUC				AUPR			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	NBI	-	-	-	-	93.5	98.1	94.8	85.1	97.5	97.6	94.6	83.8	-	-	-	-	-	-	-	-
2.	NRWRH	-	-	-	-	85.0	-	-	-	99.0	-	-	-	-	-	-	-	-	-	-	-
3.	Net CBP	-	-	-	-	-	-	-	-	-	-	-	-	82.5	80.3	82.3	83.9	-	-	-	-
4.	NMIF	-	-	-	-	-	-	-	-	-	-	-	-	83.0	82.0	82.0	80.0	81.0	78.0	74.0	71.0
5.	RWR	-	-	-	-	-	-	-	-	-	-	-	-			70.9		-	-	-	-
6.	IN-RWR/ Corank			82.2		-	-	-	-	-	-	-	-			95.1		-	-	-	-
7.	NRLMF-beta	-	-	-	-	-	-	-	-	-	-	-	-	99.0	99.0	97.5	96.4	89.7	91.3	75.5	75.5

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

TABLE XIV. Quantitative Analysis of the Deep Learning-Based Methods Used in DTI Prediction

Deep Learning based Methods																					
S. No	ML Tech.	Accuracy				Sensitivity/Recall				Precision				AUC				AUPR/MCC			
		E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N	E	IC	G	N
1.	Deep DTI			85.8				82.2				-				91.5				-	
2.	Deep DTA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			71.4	
3.	AUTO DNP	94.1	91.1	86.6	80.5	95.5	95.6	81.6	76.2	92.9	87.7	91.0	84.1	94.2	91.0	87.4	81.7	MCC			
																		88.3	82.7	73.9	61.8
4.	LASSO-DNN			81.0			-	-	-	-	-	-	-			89.0		-	-	-	-
5.	Deep Convolution DTI			75.0				85.0				70.0				80.0		-	-	-	-

E-Enzyme, IC- Ion Channel, G- G-Protein Coupled Receptor (GPCR), N- Nuclear Receptor, AUC-Area Under Curve, AUPR- Area Under Precision Recall, PPV-Positive Predicted Values, MCC- Mathew's Correlation Coefficient.

VII. Discussion

The analysis shows that the chemogenomics-based approach to DTI prediction is ideally suited to interaction prediction. A review of the qualitative and quantitative analyses offers an overview of the dataset, preprocessing, feature selection techniques, validation and ML classification techniques used in DTI prediction, all of which are discussed in this section.

A. The Dataset

The benchmark Yaminishi et al. dataset [71] is invariably used in DTI prediction, with its four enzyme (E), ion channel (IC), G-protein coupled receptor (GPCR) and nuclear receptor (NR) classes and the DTI positive pairs of each class. Apart from the benchmark dataset above, others are used as well [11], [17], [21], [26], [31]. Deep learning-based prediction works with more dynamic data. An attempt has been made in [44] to construct a negative DTI dataset, which is significant in that it facilitates the assimilation of targets not taken into the prediction process. The number of instances used, which ranges from 250 to 5500, may be increased or decreased, depending on the purpose of the research.

B. Preprocessing and Balancing Techniques

Major issues in DTI prediction are brought on by the data obtained from miscellaneous sources, which may have a different range of values or none at all. Missing values from known data are inferred, based on the observed values in the data structure. Preprocessing techniques are, generally speaking, not used on the data because they are curated when collected from different sources. When the data are incorporated, however, values may go missing or are replaced, and there is thus a need for preprocessing. The preprocessing employed in [26] to replace missing values uses the mean values of the data. Employing preprocessing techniques like data cleaning enhances the quality of the data for further processing.

From the qualitative analysis tables III-VII, it is found that the dataset used in the prediction process is unbalanced and may affect the performance of the classifiers. Balancing techniques include balancing the data using oversampling [26], [32], [33], though it increases negative outcomes. For DTI prediction, undersampling can be suggested to improve the positive outcomes.

C. Feature Extraction Methods

Feature Extraction is done to reduce the dimensionality of the input features by creating a new set of features from the original features which gains the important features of the data and also reduces the dimension of the features, which increases the speed of learning and generalization of machine learning. It can also be done through various tools available for it. In drug discovery researchers use several tools for feature extraction, the trending tools are PROFEAT and Protr for protein feature extraction, Rcpi and PADEL Descriptor for drug feature extraction. The research work which uses these tools for feature extraction are [28], [33].

D. Feature Selection Methods

Feature selection is of fundamental importance, because the extracted features increase data dimensions and result in problems with over fitting. Feature selection techniques reduce the number of features by selecting the most important ones from the given input. It is clear from the analysis that target features can be categorized into three –structural, evolutionary and sequence. While the drug feature is structural, the number of target features considered varies from 1080 to 1498. Likewise, drug features vary, depending on whether they are 1D or 2D and on the fingerprint of the drugs selected. Tables III-VII in [11]-[18] that showcase similarity-based methods only consider similarities between drug-drug, target-target and drug-target for DTI prediction, which means that only similar drugs interact with similar targets. So in similarity based methods, drug-based and target-based features are considered unimportant for DTI prediction. Further, similarity-based methods do not handle large-scale datasets. Matrix-based methods [19]-[23] consider only drug and target similarities, and no other features are taken for prediction. Also, matrix-based methods only handle small-scale datasets. Of the feature-based methods used in [25]-[33], the Sequential Forward Feature Selection (SFFS) technique is applied in [33], where the different feature sets considered are added sequentially, one by one, to evaluate the dataset. It is observed that the structural feature, which is one of the most influential target features, plays a significant role in DTI prediction, and may vary with the dataset taken. Finding the most influential features is important to feature selection. The network-based methods in [34]-[40] take different sets of features and handle them appropriately by selecting the most important drug and target features. Compact feature learning is undertaken in [39] by applying the Diffusion Component Analysis (DCA), which constructs a low-dimensional vector representation for each drug and target using diffusion distribution. It helps find the best interpretable features. The deep learning-based methods discussed [41]-[45] use the t-distributed Stochastic Neighbor Embedding (t-SNE) technique to reduce input feature dimensionality. Deep learning-based methods consider dynamic data and dynamic features. The Convolution Neural Network (CNN) used in [45] handles features with ease and finds the most potent ones. Given that deep learning-based methods deal with large-scale datasets well, future research that applies deep learning will execute DTI prediction better.

E. Validation Methods

The qualitative analysis depicts that the 10-fold Cross-Validation (CV) and 5-fold cross-validation offer better results than other CV techniques like the Leave-One-Out CV (LOOCV) and jackknife. Approaches using the LOOCV have problems with over fitting. DTI predictions are evaluated using AUC and AUPR values. The AUC values of the classifiers show better results when the 10-fold CV is used to validate the methods. AUC is chosen because it distinguishes between classes and validates the model's capacity even when the dataset is imbalanced.

F. ML Techniques Engaged in DTI Prediction

The qualitative analysis table III-VII, depicts the various classifiers used, one outclasses the rest at DTI prediction. Ranking algorithms like Bayesian ranking are used to rank DTI [20]. The SVM [19], [22] classifier, which handles target and drug features by calculating them separately and reducing prediction complexity cannot determine the relationship between the features and may produce a large number of false positives. The KNN [18], [20] falls short, performance-wise, in its inability to handle features and large-scale datasets. Ensemble learning [27] handles large-scale and high-dimensional data. The Adaboost classifier separates the data and classifies them to get the most appropriate features [32]. The decision tree manages missing data thoroughly and uses diversity to learn features based on instances for improved accuracy [33]. Logistic regression [11], [16] operates data integration strategies effectively. The DVM [29] influences features strongly in its handling of outliers. As far as feature-based methods are concerned, the random forest outperforms the rest, while the regularized least square (RLS) performs well in tandem with more influential features. In terms of performance, the WBR-DTI, VB-MK-LMF, NRLMF-beta and CNN find the best features for DTI prediction.

From the quantitative analysis table X-XIV, the progress made is evaluated using AUC values, with marked improvements in the SVM from 61.7% [19] to 96.34% [22], the KNN from 92.3% [18] to 95.4% [20], and LR from 85.1% [11] to 95.32% [16]. Among the classifiers used in DTI prediction, the SVM gives the best prediction results with an improvement of 34.64%. The random forest and decision tree used in ensemble learning give an AUC value of 90%. Adaptive Boosting and RLS give AUC values of 88.7% and 97%, respectively. The WBR-DTI and VB-MK-LMF give an AUC value of 98%, while the NRLMF-beta gives 96%.

However, the results are based on the data given as input. The new model developed may perform poorly, with imbalanced data and missing values. The qualitative analysis tables III-VII show that the dataset has more negative than positive predictions, owing to the nature of the dataset used for DTI prediction. The quantitative analysis tables X-XIV depict that matrix factorization-based methods perform best for DTI prediction, though deep learning-based methods handle large-scale data and find the most influential features and some of the papers gives light to other process like detecting adverse reaction of drugs [72]. This review has thus laid out a thorough understanding of datasets, feature selection methods and validations, as well as a comparison of the classifiers used for DTI prediction

VIII. Conclusion and Future Scope

It is concluded from the review that much research has focused chiefly on chemogenomics, and this is because DTI based on drug and target features and similarities may be found without their structures. The method works well by finding the most influential features using a range of classifiers for DTI prediction. The classifiers use only known static interaction for training the model, given that the interaction data is static. Though static data has largely been used as a benchmark dataset for interaction prediction, dynamic data may be considered so the problem of new DTI is resolved. Several studies have only considered target features (like the AAC, CTD and pseudo AAC) and the PubChem fingerprint for drugs. There are, therefore, plenty of research opportunities to predict drugs using the influence of all the features. Influential features may vary from one technique to another. There is, however, a delay in finding influential features, since one feature may not be as important for prediction as another. More data are to be considered for finding the most influential features, which is possible with the introduction of big data for prediction. The ML techniques used by the deep learning-based and matrix-based methods were found to predict DTI better than others. It is recommended, considering the above, that future researchers focus on building a negative dataset for interaction prediction. Feature scaling or feature engineering techniques may be applied to enhance the dataset. New databases can be created by collecting data from numerous sources and incorporating appropriate parameters or influential features for future research. Further, future models developed for DTI prediction must consider every feature for drug prediction. The model developed, based on ML techniques, should be able to update information on drugs and targets constantly for new interaction prediction. Thus, the model must be able to predict interaction, based on prior knowledge, without having to be trained on every occasion. Such a model is likely to offer the best interaction prediction.

References

[1] Martin L, Hutchens M, Hawkins C, Radnov A, "How much do clinical trials cost?," Nature Reviews-Drug Discovery, vol no: 16(6), pp: 381-382, June 2017, DOI: 10.1038/nrd.2017.70.

[2] Swamidass SJ, "Mining small-molecule screens to repurpose drugs," Briefing in Bioinformatics, Vol.no:12(4), pp: 327–335, 2011, DOI: 10.1093/bib/bbr028.

[3] Moriaud F, Richard SB, Adcock SA, Chanas-Martin L, Surgand JS, Ben Jelloul M, Delfaud F, "Identify drug repurposing candidates by mining the protein data bank," Briefings in Bioinformatics, vol. no: 12(4), pp: 336-340, Jul 2011, DOI: 10.1093/bib/bbr017.

[4] Chen R, Liu X, Jin S, Lin J and Liu J, "Machine learning for drug-target interaction prediction," Molecules, vol.no: 23(9), pp: 2208, 2018, DOI: 10.3390/molecules23092208.

[5] Tanaji T. Talele, Santosh A. Khedkar and Alan C. Rigby, "Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic," Current Topics in Medicinal Chemistry, vol. no: 10, pp:127, 2010, DOI: 10.2174/156802610790232251.

[6] Usha, T., Shanmugarajan, D., Goyal, A.K., Kumar, C.S. and Middha, S.K., "Recent updates on computer-aided drug discovery: time for a paradigm shift," Current topics in medicinal chemistry, vol. no: 17(30), pp.3296-3307, 2017, DOI: 10.2174/1568026618666180101163651.

[7] Jacob L, Vert J-P. "Protein-ligand interaction prediction: an improved chemogenomics approach," Bioinformatics, vol.no: 24(19), pp: 2149–2156, 2008, DOI: 10.1093/bioinformatics/btn409.

[8] Rognan D, "Chemogenomic approaches to rational drug design," British Journal of Pharmacology, vol. no: 152(1), pp: 38–52, 2007, DOI:10.1038/sj.bjp.0707307.

[9] Nath A, Kumari P, Chaube R, "Prediction of human drug targets and their interactions using machine learning methods: current and future perspectives," Methods in molecular biology, Springer, NY, USA, vol. no: 1762, pp: 21–30, 2018, doi:10.1007/978-1-4939-7756-7_2.

[10] Lü L,Zhou T, "Link prediction in complex networks: a survey", Physica A, vol. no: 390, pp: 1150–1170, 2011, Doi: 10.1016/j.physa.2010.11.027.

[11] Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R, "Combining drug and gene similarity measures for drug-target elucidation," Joural of computational biology : a journal of computational molecular cell biology , vol. no: 18(2), pp: 133–145, 2011, doi:10.1089/cmb.2010.0213.

[12] Mei J-P, Kwoh C-K, Yang P, Li XL, Zheng J, "Drug–target interaction prediction by learning from local information and neighbours," Bioinformatics, vol.no: 29(2), pp: 238–245, 2012, DOI: 10.1093/bioinformatics/bts670.

[13] Van Laarhoven T,Marchiori E, "Predicting drug–target interactions for new drug compounds using a weighted nearest neighbor profile," PloS One, vol. no: 8(6), pp: e66952, 2013, DOI: 10.1371/journal.pone.0066952.

[14] Shi J-Y, Yiu S-M, Li Y, Leung HC, Chin FY, "Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering," Methods, vol. no: 83, pp: 98–104, 2015, DOI: 10.1016/j.ymeth.2015.04.036.

[15] Buza K, "Drug–target interaction prediction with hubness aware machine learning," In: 2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI), IEEE, New York, USA, 2016, pp: 37–40, **DOI:** 10.1109/SACI.2016.7507416.

[16] Zhang W, Chen Y, Li D, "Drug–target interaction prediction through label propagation with linear neighborhood information," Molecules, vol. no: 22(12), pp: 2056, 2017, DOI: 10.3390/molecules22122056.

[17] Zhang X, Li L, Ng MK, Zhang S, "Drug–target interaction prediction by integrating multiview network data," Computational Biology and Chemistry, vol. no: 69, pp: 185–193, 2017, DOI: 10.1016/j.compbiolchem.2017.03.011.

[18] Shi Z, Li J, "Drug–target interaction prediction with weighted Bayesian ranking," In: Proceedings of the 2nd International Conference on Biomedical Engineering and Bioinformatics, ACM, London, United Kingdom, 2018, pp: 19–24.

[19] Rendle S, Freudenthaler C, Gantner Z, Zeno Gartner, Lars Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, McGill, Canada, 2009, pp: 452–461, DOI: 10.1145/3278198.3278210.

[20] Gönen M, "Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization," Bioinformatics, vol. no: 28(18), pp: 2304–2310, 2012, DOI:10.1093/bioinformatics/bts360.

[21] Li L, Cai M, "Drug target prediction by multi-view low rank embedding," IEEE/ACM Transactions on Computational Biology and Bioinformatics vol. 16, no.5, pp.1712-1721, 1 Sep-Oct 2019, DOI: 10.1109/TCBB.2017.2706267.

[22] Bolgár B, Antal P, "VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization," BMC Bioinformatics, vol. no: 18(1), pp: 440, 2017, DOI: 10.1186/s12859-017-1845-z.

[23] Huang YA, You ZH, Chen X, "A Systematic Prediction of Drug-Target Interactions Using Molecular Fingerprints and Protein Sequences," Current protein & peptide science, vol. no: 19(5), pp: 468-478, 2018, DOI: 10.2174/1389203718666161122103057.

[24] M. Caro-Martínez, G. Jiménez-Díaz, J. A. Recio-García. "Local Model-Agnostic Explanations for Black-box Recommender Systems Using Interaction Graphs and Link Prediction Technique", International Journal of Interactive Multimedia and Artificial Intelligence, 2021, DOI: 10.9781/ijimai.2021.12.001.

[25] Van Laarhoven T, Nabuurs SB, Marchiori E, "Gaussian interaction profile kernels for predicting drug–target interaction," Bioinformatics, vol. no: 27(21), pp: 3036–3043, 2011, DOI: 10.1093/bioinformatics/btr500.

[26] Ezzat A, Wu M, Li X-Li, Kwoh Chee-Keong, " Drug–target interaction prediction via class imbalance-aware ensemble learning," BMC Bioinformatics, vol. no: 17(19), pp: 509, 2016, DOI:10.1186/s12859-016-1377-y.

[27] Nascimento A.C., Prudêncio R.B., Costa I.G., "A multiple kernel learning algorithm for drug–target interaction prediction," BMC Bioinformatics, vol. no:17, pp:46 2016, DOI: 10.1186/s12859-016-0890-3.

[28] Lan W, Wang J, Li M, Liu J, Li Y, et al., "Predicting drug–target interaction using positive - unlabeled learning," Neurocomputing, vol. no: 206, pp: 50–57, 2016, DOI: 10.1016/j.neucom.2016.03.080.

[29] Li Z, Han P, You Z-H, Li X, Zhang Y, Yu H, et al., "In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences," Scientific Reports, vol. no: 7(1), pp: 11174, 2017, DOI: 10.1038/s41598-017-10724-0.

[30] Ohue M, Yamazaki T, Ban T, Akiayama Y, "Link mining for kernel based compound–protein interaction predictions using a chemogenomics approach," In: International Conference on Intelligent Computing, Springer, Cham, Switzerland, 2017, pp: 549–558, DOI: 10.1007/978-3-319-63312-1_48.

[31] Zhang J, Zhu M, Chen P, Wang B, "DrugRPE: random projection ensemble approach to drug–target interaction prediction," Neurocomputing, vol. no: 228, pp: 256–262, 2017, DOI: 10.1016/j.neucom.2016.10.039.

[32] Rayhan F, Ahmed S, Shatabda S, et al., "iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting," Scientific Reports, vol. no: 7(1), pp: 17731, 2017, DOI:10.1038/s41598-017-18025-2.

[33] Sharma A, Rani R, "BE-DTI: ensemble framework for drug target interaction prediction using dimensionality reduction and active learning," Computer Methods and Programs in Biomedicine, vol. no: 165, pp: 151–162, 2018, DOI:10.1016/j.cmpb.2018.08.011.

[34] Cheng F, Liu C, Jiang J, et al., "Prediction of drug–target interactions and drug repositioning via network-based inference," PLOS Computational Biology, vol. no: 8(5), pp: e10025032012, 2012, DOI: 10.1371/journal.pcbi.1002503.

[35] Chen X, Liu M-X, Yan G-Y, "Drug–target interaction prediction by random walk on the heterogeneous network," Mol Biosyst, vol. no: 8(7), pp: 1970–1978, 2012, DOI: 10.1039/C2M00002D.

[36] Chen H, Zhang Z, "A semi-supervised method for drug–target interaction prediction with consistency in networks," PloS One, vol. no: 8(5), pp: e62975, 2013, DOI: 10.1371/joural.poe.0062975.

[37] Peng L, Liao B, Zhu W, Li Z, Li K, "Predicting drug–target interactions with multi-information fusion," IEEE J Biomed Health Inform, vol. no: 21(2), pp: 561–572, 2015, DOI: 10.1109/JBHI.2015.2513200.

[38] Seal A, Ahn Y-Y, Wild DJ, "Optimizing drug–target interaction prediction based on random walk on heterogeneous networks," J Chem, vol. no: 7(1), pp: 40, 2015, DOI: 10.1186/s13321-015-0089-z.

[39] Huang Y, Zhu L, Tan H, et al., "Predicting drug-target on heterogeneous

network with co-rank," In: International Conference on Computer Engineering and Networks, Springer, Cham, Switzerland, 2018, 571–581, DOI: 10.1007/978-3-030-14680-1_63.

[40] Ban T, Ohue M, Akiyama Y, "NRLMFβ: beta-distribution rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction," Biochemistry and Biophysics Reports, vol. no: 18, pp: 100615, 2019, DOI: 10.1016/j.bbrep.2019.01.008.

[41] Wen M, Zhang Z, Niu S, et al, "Deep-learning-based drug– target interaction prediction," J Proteome Res, vol. no: 16(4), pp: 1401–1409, 2017, DOI:1 0.1186/s12911-020-1052-0 .

[42] Öztürk H, Özgür A, Ozkirimli E, "DeepDTA: deep drug–target binding affinity prediction," Bioinformatics, vol. no: 34(17), pp: i821–i829, 2018, DOI: 10.1093/bioinformatics/bty593.

[43] Wang L, You Z-H, Chen X, et al, "A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network," Journal of Computational Biology: Journal of Computational molecular biology, vol. no: 25(3), pp: 361–373, 2018, DOI: 10.1089/cmb.2017.0135.

[44] You J, McLeod RD, Hu P, "Predicting drug–target interaction network using deep learning model," Computational Biology Chemistry, vol. no: 80, pp: 90–101, 2019, DOI: 10.1016/j.compbiolchem.2019.03.016.

[45] Lee I, Keum J, Nam H, "DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences," PLoS Computational Biology, vol. no: 15(6), pp: e1007129, 2019, DOI: 10.1371/journal.pcbi.1007129.

[46] Kanehisa M, Araki M, Goto S, et al., "KEGG for linking genomes to life and the environment," Nucleic Acids Research, vol. no: 36, pp: D480–484, 2007, DOI: 10.1093/nar/gkm882.

[47] Kanehisa M, Goto S, Hattori M, M Araki, M Hirakawa, "From genomics to chemical genomics: new developments in KEGG," Nucleic Acids Research, vol. no: 34, pp: D354–D357, 2006, DOI: 10.1093/nar/gkj102.

[48] Gaulton A, Hersey A, Nowotka M, et al., "The ChEMBL database in 2017," Nucleic Acids Research, vol. no : 45(D1), pp: D945–954, 2016, DOI: 10.1093/nar/gkw1074.

[49] Kringelum J, Kjaerulff SK, Brunak S, et al., "ChemProt-3.0: a global chemical biology diseases mapping," Database: the journal of biology databases and curation, vol. 2016 bav123, 2016, DOI: 10.1093/database/bav123.

[50] Wagner AH, Coffman AC, Ainscough BJ, et al, "DGIdb 2.0: mining clinically relevant drug–gene interactions," Nucleic Acids Research, vol. no: 44(D1), pp: D1036–1044, 2016, DOI: 10.1093/nar/gkv1165 .

[51] Wishart DS, Feunang YD, Guo AC, et al., "Drugbank 5.0: a major update to the drugbank database for 2018," Nucleic Acids Research, vol. no: 46(D1), pp: D1074–1082, 2017, DOI: 10.1093/nar/gkx1037.

[52] Kanehisa M, Furumichi M, Tanabe M, et al., "KEGG: new perspectives on genomes, pathways, diseases and drugs," Nucleic Acids Research, vol. no: 45(D1), pp: D353–361, 2016, DOI: 10.1093/nar/gkw1092.

[53] HMS_LINCS: LINCS Pilot Phase Joint Project: Sensitivity measures of six breast cancer cell lines to a library of small molecule kinase inhibitors (drug combination treatments). Dataset 2 of 2: Mean cell count and mean normalized growth rate inhibition values across technical replicates., 2016.

[54] Von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R, "PROMISCUOUS: a database for network-based drug-repositioning," Nucleic Acids Research, Vol no: 36, Jan 2011, DOI:10.1093/nar/gkq1037.

[55] Szklarczyk D, Santos A, von Mering C, et al., "STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data," Nucleic Acids Research, vol. no: 44(D1), pp: D380–384, 2015, DOI: 10.1093/nar/gkv1277.

[56] Günther S, Kuhn M, Dunkel M, et al., "Supertarget and matador: resources for exploring drug-target relationships," Nucleic Acids Research, vol. no: 36, pp: D919–922, 2008, DOI: 10.1093/nar/gkm862.

[57] Chen X, Ji ZL, Chen YZ, "TTD: therapeutic target database," Nucleic Acids Research, vol. no: 30(1), pp: 412–415, 2002, DOI: 10.1093/nar/30.1.412.

[58] Jeske L, Placzek S, Schomburg I, et al., "Brenda in 2019: a European ELIXIR core data resource," Nucleic Acids Research, vol. no: 47(D1), pp: D542–549, 2019, DOI: 10.1093/nar/gky1048.

[59] Ursu O, Holmes J, Bologa CG, et al., "DrugCentral 2018: an update," Nucleic Acids Research, vol. no: 47(D1), pp: D963–970, 2018, DOI: 10.1093/nar/gky963.

[60] Wang C, Hu G, Wang K, et al., "PDID: database of molecular level putative protein–drug interactions in the structural human proteome," Bioinformatics, vol. no: 32(4), pp: 579–586, 2016, DOI: 10.1093/bioinformatics/btv597.

[61] Nguyen D-T, Mathias S, Bologa C, et al., "Pharos: collating protein information to shed light on the druggable genome", Nucleic Acids Research, vol. no: 45(D1), pp: D995–D1002, 2017, DOI: 10.1093/nar/gkw1072.

[62] Kim S, Thiessen PA, Bolton EE, et al., "PubChem substance and compound databases", Nucleic Acids Research, vol.no: vol.no: 44(D1), pp: D1202–1213, 2016, DOI: 10.1093/nar/gkv951.

[63] Siramshetty VB, Eckert OA, Gohlke B-O, et al, "SuperDRUG2: a one stop resource for approved/marketed drugs", Nucleic Acids Research, vol. no: 46(D1), pp: D1137–1143, 2018, DOI: 10.1093/nar/gkx1088.

[64] Fang, H., Su, Z., Wang, Y., Miller, A., Liu, Z., Howard, P. C., Tong, W., & Lin, S. M, "Exploring the FDA adverse event reporting system to generate hypotheses for monitoring of disease characteristics", Clinical pharmacology and therapeutics, vol. no: 95(5), pp: 496–498, 2014, DOI: 10.1038/clpt.2014.17.

[65] Kuhn M, Letunic I, Jensen LJ, Bork P, "The SIDER database of drugs and side effects," Nucleic Acid Research, vol. no: 44(D1), pp: D1075-1079, 2015, DOI: 10.1093/nar/gkv1075.

[66] Pawson AJ, Sharman JL, Benson HE, et al., "The IUPHAR/BPS guide to pharmacology: an expert-driven knowledgebase of drug targets and their ligands," Nucleic Acids Research, vol. no: 42(D1), pp: D1098–1106, 2013, DOI: 10.1093/nar/gkt1143.

[67] Kumar, R., Chaudhary, K., Gupta, S. et al., "CancerDR: Cancer Drug Resistance Database," Scientific Reports, vol. no:3, pp: 1445, 2013, DOI:10.1038/srep01445.

[68] GilsonMK, Liu T, BaitalukM, et al., "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," Nucleic Acids Research, vol. no: 44(D1), pp: D1045–1053, 2016, DOI: 10.1093/nar/gkv1072.

[69] Sterling and Irwin, J. "ZINC- database for ligand discovery which has all the details about chemical matter for biological targets," Journal of Chemical Information and modelling, vol. no: 55(11), pp: 2324-2337, 2015, doi = 10.1021/acs.jcim.5b00559.

[70] Roth BL, Kroeze WK, Patel S, Lopez E, "PDSP Ki -The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrassment of riches?," The Neuroscientist, vol. no: 6, pp: 252–262, 2000, DOI:10.1177/107385840000600408.

[71] Yamanishi Y, Kotera M, Kanehisa M, Goto S, "Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," Bioinformatics, vol. no: 26(12), pp: i246–i254, 2010, DOI: 10.1093/bioinformatics/btq176.

[72] Carrasco, Rafael San Miguel, "Detection of Adverse Reaction to Drugs in Elderly Patients through Predictive Modeling." International Journal of Interactive Multimedia and Artificial Intelligence, vol.no:3(6), pp:52-56, 2016, DOI:10.9781/ijimai.2016.368.

A. Suruliandi

Dr. A. Suruliandi received the B.E., degree in electronics and communication engineering from the Coimbatore institute of Technology, Coimbatore, India, in 1987. He completed his M.E., degree in computer science and engineering from the Government College of Engineering, Tirunelveli, India, in 2000. And he pursued his Ph.D degree from Manonmaniam Sundaranar University, Tirunelveli, in 2009. He is currently working as a professor with the Department of Computer Science and Engineering, Manonmaniam Sundaranar University. He has more than 29 years of experience in teaching. He has been author of 50 articles in international journals, 23 articles in IEEE Xplore publications, 33 in national conferences, and 13 in international conferences. His interested research areas are remote sensing, image processing, and pattern recognition.

T. Idhaya

T. Idhaya received M.Sc., degree in Computer Science from St. Xavier's College (Autonomous), Tirunelveli, India, in 2016. She has completed her M.phil degree in Manonmaniam Sundaranar University, Tirunelveli, India, in 2017. She is currently pursuing her Ph.D degree in Manonmaniam Sundaranar University, Tirunelveli, India. Her area of interest is Image processing, Machine learning and Big data.

S. P. Raja

S. P. Raja is born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. Currently he is working as an Associate Professor in the School of Computer Science and Engineering in Vellore Institute of Technology, Vellore, Tamilnadu, India. He published 75 papers in International Journals, 24 in International conferences and 12 in national conferences. Dr. Raja is an Associate Editor of the Journal of Circuits, Systems and Computers, Computing and Informatics, International Journal of Interactive Multimedia and Artificial Intelligence, Brazilian Archives of Biology and Technology, International Journal of Image and Graphics, and International Journal of Biometrics.