Briefings in Bioinformatics, 00(00), 2021, 1–20

doi: 10.1093/bib/bbab046
Case Study

PreDTIs: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques

S. M. Hasan Mahmud, Wenyu Chen, Yongsheng Liu, Md. Abdul Awal, Kawsar Ahmed, Md. Habibur Rahman and Mohammad Ali Moni

Corresponding authors: Mohammad Ali Moni, WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, UNSW Sydney, Australia. Email: m.moni@unsw.edu.au; Wenyu Chen, PhD, Computational Intelligence Laboratory (CIL), School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. Email: cwy@uestc.edu.cn

Abstract

Discovering drug–target (protein) interactions (DTIs) is of great significance for researching and developing novel drugs, having a tremendous advantage to pharmaceutical industries and patients. However, the prediction of DTIs using wet-lab experimental methods is generally expensive and time-consuming. Therefore, different machine learning-based methods have been developed for this purpose, but there are still substantial unknown interactions needed to discover. Furthermore,

Wenyu Chen, PhD, Computational Intelligence Laboratory (CIL), School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

S. M. Hasan Mahmud received the PhD degree in computer science from the University of Electronic Science and Technology of China, China. Since 2013, he has been serving as a Faculty Member with the Department of Software Engineering, Daffodil International University, Bangladesh. He has published several conference and journal papers. His research interests include machine learning, deep learning, bioinformatics, drug discovery and pattern recognition.

Wenyu Chen received the PhD degree in computer science from the University of Electronic Science and Technology of China, China, where he is currently a Professor with the School of Computer Science and Engineering. His research interests include pattern recognition, bioinformatics, natural language processing and neural networks

Yongsheng Liu received the PhD degree in computer science and technology with the Department of Computer Science and Engineering, University of Electronic Science and Technology of China, China. His current research interests include neural networks, intelligent computation, deep learning and optimization.

Md. Abdul Awal received his PhD in Biomedical Engineering from The University of Queensland, Australia in 2018. He is now serving as an Associate Professor in the ECE Discipline, Khulna University. His research interest includes signal processing, especially, biomedical signal processing, big data analysis, image processing, time-frequency analysis, machine learning algorithms, deep learning, optimization, expert system with applications, computational intelligence in biomedical engineering.

Kawsar Ahmed received his BSc and MSc Engineering Degree in Information and Communication Technology (ICT) at Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. He has achieved gold medals for engineering faculty first both in BSc (Engg.) and MSc (Engg.) degree from the university for his academic excellence. Currently, he is serving as an Assistant Professor at the same department. He has more than 200 publications in IEEE, IET, OSA, Elsevier, Springer, ISI and PubMed indexed journals. He has published two books on bioinformatics and photonic sensor design. He is research coordinator of Group of Biophotomatiχ. He is also member of IEEE, SPIE and OSA. He holds top position at his department as well as university from 2017 to 2019 and 1st, 2nd and 4th top researcher (Scopus indexed based) in Bangladesh, 2019 to 2017, respectively. His research interests include sensor design, bio-photonics, nanotechnology, data mining and bioinformatics.

Md Habibur Rahman received PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is an Assistant Professor in the Department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. His research interest encompasses artificial intelligence, machine learning, pattern recognition, medical image processing and clinical bioinformatics.

Mohammad Ali Moni is a Research Fellow and Conjoint Lecturer at the University of New South Wales, Australia. He received his PhD degree in Clinical Bioinformatics and Machine Learning from the University of Cambridge. His research interest encompasses artificial intelligence, machine learning, data science and clinical bioinformatics

Submitted: 13 October 2020; **Received (in revised form):** 25 January 2021

data imbalance and feature dimensionality problems are a critical challenge in drug-target datasets, which can decrease the classifier performances that have not been significantly addressed yet. This paper proposed a novel drug–target interaction prediction method called PreDTIs. First, the feature vectors of the protein sequence are extracted by the pseudo-position-specific scoring matrix (PsePSSM), dipeptide composition (DC) and pseudo amino acid composition (PseAAC); and the drug is encoded with MACCS substructure fingerings. Besides, we propose a FastUS algorithm to handle the class imbalance problem and also develop a MoIFS algorithm to remove the irrelevant and redundant features for getting the best optimal features. Finally, balanced and optimal features are provided to the LightGBM Classifier to identify DTIs, and the 5-fold CV validation test method was applied to evaluate the prediction ability of the proposed method. Prediction results indicate that the proposed model PreDTIs is significantly superior to other existing methods in predicting DTIs, and our model could be used to discover new drugs for unknown disorders or infections, such as for the coronavirus disease 2019 using existing drugs compounds and severe acute respiratory syndrome coronavirus 2 protein sequences.

Key words: drug–target interaction; drug chemical structure; protein sequence; data imbalance; feature selection; SARS-CoV-2

Introduction

Prediction of new drug–target (protein) interactions (DTIs) is a fundamental stage in the drug development and drug discovery pipeline [1–3]. Drug repurposing is a growing trend in pharmaceutical science for drug discovery giving emphasis on identifying the unknown interactions between existing drugs and new target proteins. The development of the human genome and the expansion of the molecular medicine project are useful to predict the new target of drugs. In the past years, lots of efforts have been imposed on discovering unknown drugs, but very few new drugs got approvals by the Food and Drug Administrations (FDA) and reached people [4], whereas a huge number of drugs got a rejection in the clinical tests because of the unacceptable toxicity [5]. The wet-lab experiments of DTIs are usually time-consuming, labor-intensive and costly [6]; therefore, such failures are not easy to accept and wasted a lot of money. Generally, the cost of novel drugs is about $1.8 billion, and it takes approximately 13 years to develop [7]. Therefore, researchers are highly motivated to build machine learning (ML)-based techniques to detect DTIs, which can successfully reduce the search space of the drug–target candidates to be examined by wet-lab experiments to minimize the effort and cost. Recently, ML-based computational methods become more beneficial due to large heterogeneous drugs and protein data. Various online databases are available to use for the application in the prediction of the drug–target interactions (DTIs) [8], such as KEGG [9, 10], DrugBank [11], TTD [12,13] and STITCH [14].

Different ML methods have recently been applied to predict DTIs based on the various types of datasets [15, 16]. The computational drug–target methods can be divided into three groups: ligand-based methods [17,18], docking-based methods [19,20] and chemogenomic methods [21]. The ligand-based methods discover DTIs based on the similarity of the proteins' ligands. The docking-based methods exploit three-dimensional (3D) structure data of a protein and then execute models to evaluate the prediction probability that it binds with a specific drug through binding affinity and energy. Finally, chemogenomic methods [22] generally use the genomic and chemical information of target proteins and drugs. However, the first and second categories methods have some limitations due to the insufficiently known ligands and 3D structures of proteins. Therefore, chemogenomics methods are more popular to identify DTIs. The prediction task of the chemogenomic model could be solved by using sophisticated ML algorithms [23]. Here, the model takes known interaction data together with information (chemical and sequence) of the drugs–proteins involved to train the algorithm and consequently detect new interactions from the trained algorithm. According to the current review paper [24], the ML methods can be classified as similarity-based approaches or feature-based approaches [21,22,25]. Similarity-based approaches include matrix factorization [27–30], kernel-based approaches and graph-based approaches [31]. Different ML classifiers [32] such as XGBoost [33], deep learning [34], SVM [35], fuzzy logic [36,37] and nearest neighbor [1] have been effectively applied on these types of prediction purposes. Whereas feature-based approaches consider drug chemical and protein sequence feature vectors as input and represent the class label by binary value (1 and 0), indicating the interacted and non-interacted pairs in the datasets. These methods can discover possible interactions from features that are more effective. Yamanishi *et al.* [38] combined chemical structures of drug compounds and sequences information of target proteins in a unified space to contract the drug-protein interactions network. Furthermore, the same author proposed an integrated model to detect the pharmacological similarity of molecular compounds in a bipartite graph (BG) inference [39]. This model combined chemical and pharmacological information with the topological network and used a distance learning classifier to train the model. The primary purpose of this model was to discover potential similar DTIs using pharmacological effect. Based on the similarity-based network of genomics data, Hao *et al.* [40] introduced a robust regularized least squares with kernel fusion (RLS-KF) classifier that utilizes nonlinear kernel fusion (KF) technique with different kernel matrices to obtain the shared and complementary information for identifying new DTIs. Another approach, called KBMF2K, was introduced in [27] that integrates dimensionality reduction (DR), matrix factorization and binary classifiers. This method provides full derivation using variational approximation.

Li *et al.* [41] extracted the PSSM features of proteins from the amino acid sequence and the drug chemical structure transferred to substructure fingerprints, where a discriminative vector-based ML algorithm is developed. To minimize feature dimensionality, principal component analysis (PCA) is applied to make features into the low-dimensional space for protein and drugs. Finally, the local binary pattern (LBP) technique is used to evaluate the LBP histogram from the reduced features. Rayhan *et al.* [42] developed a boosting classifier-based method to predict potential DTIs from gold standard datasets using evolutionary information and structural features of target proteins. In this

work, the authors proposed a cluster undersampling technique (CUS) to handle the data imbalance problem. Similarly, Wang *et al*. [43] employed a rotation forest (RF) classifier with an auto-covariance method to detect interactions from PSSM (protein features) and fingerprint (drug features) vectors. In another work, a DNN-deep learning method is proposed in [44], a LASSO-random forest approach is introduced in [45] and an ensemble method with a random projection is proposed in [46]. Recently, our previous work [47] presented XGBoost based approach with sequence, evolutionary and structural information to identify DTIs. In this paper, a cluster-based undersampling technique was proposed to manage the data imbalance issue, and also a new feature selection method was developed to select the optimal set of features. In another work [26], structural features extracted from protein sequences, an oversampling-SMOTE, are independently employed for balancing the drug-target datasets. During the prediction of DTIs, the high-dimensionality of features is a complex issue. Therefore, DR is an essential part of the prediction task. There are different types of DR techniques, such as linear discriminant analysis (LDA) [48], PCA [49], genetic algorithm (GA) [50] and relief [51] have been applied to select the suitable features from the datasets for accurate prediction. Considering the advantages of related approaches, more related approaches have also been proposed. Thafar *et al*. [52] introduced a method DTiGEMS+ that predicts DTIs using graph mining graph embedding and similarity techniques. Their method combines feature-based and similarity-based approaches and models the prediction of potential DTIs as a link identification problem in a heterogeneous network. DTiGEMS+ uses the heterogeneous network by enhancing the positive DTIs graph using two more matching graphs: target–target similarity and drug–drug similarity. DTiGEMS+ incorporates multiple target–target similarities and drug–drug similarities into a heterogeneous graph after utilizing a similarity selection technique and a fusion algorithm. In most recent studies, ML methods, similarity metrics and handcrafted features have been proposed to discover DTIs. Manoochehr *et al*. [53] claimed that these approaches could not learn the fundamental associations between drugs and targets. Therefore, they proposed a novel framework for identifying DTIs that acquire latent features from DTI network. Another recent work was introduced in [54] that uses a deep-walk embedding concept to predict DTIs from a molecular association's network. This network is constructed by combining the associations among protein, drug, disease, micro RNA and long non-coding RNA (lncRNA). Moreover, we summarized some of the exiting methods related to DTI in Table 1.

Since wet-lab experiments for predicting DTIs are expensive ($1.8 billion), time-consuming (around a decade) and laborious, it completely motivated us to develop ML-based methods for discovering potential interactions efficiently. However, establishing such ML approaches is not easy, but it is urgently needed, as existing approaches that identify potential DTIs suffer from high false-positive rates. Still, vast numbers of possible DTIs interactions are undiscovered, which is also necessary to predict and help develop new drugs. Therefore, the novelty and contributions of this research include: (i) predict the novel DTIs from drug chemical structure and protein sequence; (ii) utilize the multi-feature fusion for predicting novel DTIs; (iii) propose a data balancing algorithm address to handle the imbalance issue in datasets that did not effectively address in the existing approaches; (iv) develop an effective feature selection algorithm to remove the redundant and noisy information and (v) provide satisfactory prediction performance for all the four benchmark datasets.

In this study, we proposed a new ML-based model called PreDTIs for DTIs prediction. Firstly, we use MACCS fingerprints, PsePSSM, PseAAC and DC to extract the drug chemical structure features and protein sequence features. Then these three types of protein features integrate with drug features to make drug–target datasets for accurate DTIs prediction. Secondly, since the drug–target dataset is highly imbalanced, we propose a new undersampling technique to manage the imbalance issue of positive and negative datasets. Thirdly, Modified Incremental Feature Selection (MoIFS) develops to select the optimal features, removing noisy and redundant features and providing potential features for accurate prediction. Finally, after comparing different ML classifiers, the LightGBM classifier is selected for predicting DTIs from balanced and selected features. Five-fold CV test is carried out on the datasets; different parameters are chosen for the features to fix the best settings of the model. Average area under the curve (AUC) values of PreDTIs on enzyme (EN), ion channel (IC), GPCR and NR are 0.9656, 0.9612, 0.9249 and 0.8652, respectively. The results indicate that our proposed model significantly improved the prediction performance of DTIs compared to the other existing methods.

The rest part of the paper is structured as follows: Materials and methods section describes the detail of the gold standard datasets, feature extraction, data balancing, feature selection, and classifiers we employed in this paper. In the Results and Discussion section, performances and experimental results are provided, and a comparison with the literature is made. In the next section, our proposed model that can help discover new drugs to treat coronavirus disease 2019 (COVID-19) infection is described, and finally, a discussion and conclusion are drawn in this paper.

Materials and methods

A schematic diagram of our proposed PreDTIs model is shown in Figure 1. At first, drug chemical structures (SMILE format) and protein sequences (FASTA format) are collected from DrugBank and KEGG databases using their specific access IDs. Different feature generation methods were applied to the drug and protein sequences to generate a variety of features. Afterwards, balancing techniques are used on extracted features to manage the datasets' imbalance issue and drug–target features are selected through the newly developed feature selection technique to boost prediction performances. Finally, the model is trained using LightGBM classifier on the reduced features.

Drug–target datasets

In our research, we use four types of protein targets gold standard datasets also known as benchmark datasets, i.e. EN, IC, GPCR and nuclear receptor (NR) released by Yamanishi *et al*. [38]. Mainly, only drug IDs and protein IDs are considered from their datasets. After that, we collected the drug chemical structures and protein sequences of these four types of datasets from the DrugBank [55], SuperTarget and Matador [56], KEGG BRITE [57] and BRENDA [58] databases. After counting them, the number of known interaction (positive samples) pairs in each dataset is 2926, 1476, 635 and 90, respectively. Finally, the total number of 5127 known interacted pairs was found. The detailed information about the drug–target datasets is shown in Table 2. Note that these gold standard datasets have been exploited in recent various state-of-the-art methods [41,42,47,59] by researchers.

Table 1. The existing DTIs prediction methods

Authors	Datasets	Methodology/techniques	Cross-validation	Performance parameters
Yuan *et al.* [50]	DrugBank	Classifier: ensemble learning and LambdaMART, feature extraction: GD (general descriptors-RDKit), composition, transition and distribution (CTD-PROFEAT)	5-fold CV	Area under precision-recall curve (AUPR)
Li *et al.* [51]	Gold standard dataset	Classifier: discriminative vector machine, feature extraction: PSSM and local binary pattern and fingerprint, feature dimension reduction: PCA	5-fold CV	Pre, ACC, SE, MCC, AUC
Meng *et al.* [52]	Gold standard dataset	Classifier: RVM feature extraction: BIGP, PSSM feature dimension reduction: PCA	5-fold CV	Ac, SE, Precision, MCC
Luo *et al.* [53]	DrugBank, HPRD, comparative toxicogenomics, SIDER	Method: vector space projection scheme (network integration pipeline)	10-fold CV	AUPR
He *et al.* [54]	Davis, Metz, KIBA	Method: SimBoost and SimBoostQuant	5-fold CV	RMSE, AUC, AUPR, CI
Laarhoven *et al.* [55]	Gold standard dataset	Classifier: RLS, feature generation: similarity matrices	10-fold CV	AUC
Buza *et al.* [51]	Gold standard dataset, kinase	Method: projection-based ensemble-BLMs, ECkNN	5-fold CV	AUC, AUPR
Kuang *et al.* [56]	DrugBank	Method: RLS (regularized least squares) and semi-supervised link prediction	10-fold CV	AUC, AUPR
Cheng *et al.* [57]	Gold standard dataset, DrugBank	Method: three supervised similarity network-based inference	10-fold CV	AUC

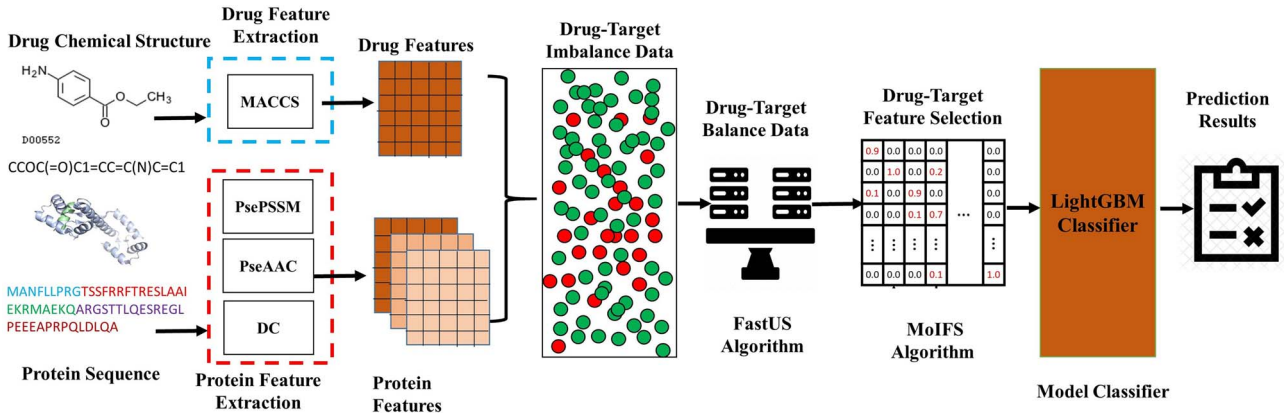

Figure 1. The workflow of identifying DTIs from the chemical structure of drug compounds and the target sequence of proteins.

Table 2. Statistics of the dataset used in this study

Datasets	Drugs	Targets	Interaction
Enzyme	445	664	2926
Ion Channel	210	204	1476
GPCR	223	95	635
NR	54	26	90

Generally, the DTIs network is visualized by a BG, where nodes of the graph represent the drugs or proteins, and the edges indicate the known interactions between these nodes (drugs and targets). Most importantly, this BG holds a small number of edges. For example, EN has $445 \times 664 = 295\,480$ edges in the BG, and only 2926 edges are known interactions (positive samples). Therefore, the possible $295\,480 - 2926 = 292\,554$ unknown interactions (negative samples) are greater than the known interactions, creating a major biasing problem. To solve the bias caused by the imbalanced data, we develop a new FastUS algorithm to balance the negative samples with the same number of positive samples (e.g. EN: 2926 positive/2926 negative) to assess classification performance.

Feature extraction methods

Drug features

Different types of software have been developed as a descriptor to calculate and represent the drug compounds in the past few

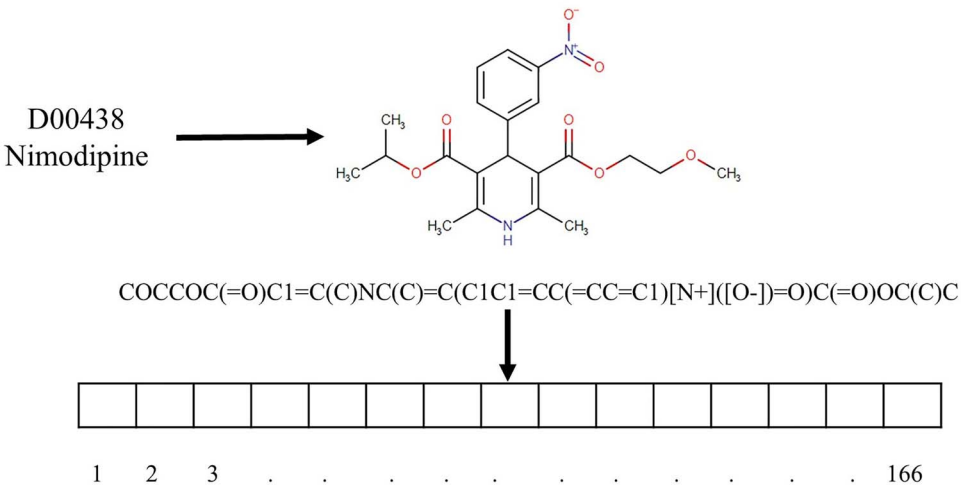

Figure 2. Schematic representation of a drug molecular substructure pattern fingerprint.

years. But, recent studies indicate that molecular substructure fingerprints (MSFs) can efficiently represent drug chemical structure under consideration [26,35,60,61]. MSFs are string representations of drug chemical structures aimed to improve chemical database searching and analysis efficiency. They can encode the 2D and 3D features of drug molecules. Among the various types of fingerprints, MSF performs well for small molecules such as drugs, while atom-pair fingerprints are the best for large molecules such as peptides. MSF directly extracts molecular structure in binary bits, the presence (one-1) or absence (zero-0) of specific sub-structures in the drug molecule. It represents a molecule into large fragments. It can retain the whole complexity of drug molecules and thereby not generate any error features from the molecular structures. Most importantly, the process provides a complete relationship between molecular property and structure. Thus, a molecule is represented as a Boolean array and described according to fingerprints of structural keys. Here, SMARTS (predefined dictionary of substructure patterns) pattern and fingerprint bit have a one-to-one relation. In the SMARTS pattern, if the substructure is present in the drug molecule, the fingerprint bit is set to one (1); otherwise, it is set to zero (0) if the substructure is absent. As an example, a substructure fingerprint dictionary for a drug molecule is shown in Figure 2.

In this experiment, we used the MACCS (Molecular ACCess System) fingerprint to create the substructure dictionary. The fingerprint of MACCS applies a dictionary of MDL keys, and there are two sets of MACCS keys (one with 960 keys and the other containing a subset of 166 keys); only the shorter fragment definitions are available to the public. These 166-D features are generated for each drug structure in this experiment. Full development information of MACCS fingerprints can get from OpenBabel and all the processes performed on the ChemoPy [62] software package.

Target features

Pseudo position specific scoring matrix (PsePSSM). To represent the character of the amino acid (AA) sequence correctly, we used PsePSSM to extract the sequence and evolution information of the target protein sequence, which is introduced by Chou and Shen [63]. This technique is extensively applied in bioinformatics research because it not only encodes sequence information from protein sequence; it also reflects evolutionary information as well.

If a protein sequence with L AA residues, PSSM is used as its descriptor proposed by Jones *et al.* [64]. For a protein sequence, a PSSM can be expressed as follows:

$$M_{\mathrm{PSSM}}=\begin{bmatrix} P_{1\to1} & P_{1\to2} & \cdots & P_{1\to j} & \cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to2} & \cdots & P_{2\to j} & \cdots & P_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ P_{i\to1} & P_{i\to2} & \cdots & P_{i\to j} & \cdots & P_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ P_{L\to1} & P_{L\to2} & \cdots & P_{L\to j} & \cdots & P_{L\to20} \end{bmatrix}, \tag{1}$$

where $P_{i,j}$ represents the residue score of i-th in the AA sequence being substituted to the j-th AA residue, which is searched by the PSI-BLAST [65] tool through the Swiss-Prot database to generate PSSM on a server machine. In this experiment, we set the PSI-BLAST parameters: the number of iterations is three, the threshold E-value is 0.001, and the remaining parameters are left as default. The indicators in PSSM are regularized using the following Equation (2).

$$f(x)=\frac{1}{1+e^{-x}} \tag{2}$$

A protein sequence can be formulated as $L \times 20$ matrix and the length of AA in the inputted protein sequence is different; therefore, PSSM matrics with different length need to transform into the same dimension using the following Equations (3) and (4):

$$M_{\mathrm{PSSM}}=[P_1,P_2,\ldots,P_{20}] \tag{3}$$

$$P_j=\frac{1}{L}\sum_{i=1}^{L}P_{ij}\,(j=1,2,\ldots,20), \tag{4}$$

where P_j is the average score of each target protein. Here, only the score of the AA residue of the i-th position being substituted by the j-th AA residue not considers any sequence information of target protein sequence. To overcome the limitation, PsePSSM is used in this study. Therefore, the full form of feature extraction produce is:

$$M^{\xi}_{\mathrm{PsePSSM}}=[\alpha^{\xi}_1,\alpha^{\xi}_2,\alpha^{\xi}_3,\ldots,\alpha^{\xi}_j,\ldots,\alpha^{\xi}_{20}] \tag{5}$$

$$\alpha_1^{\xi} = \frac{1}{L-\xi}\sum_{i=1}^{L-\xi}\left[P_{i,j} - P_{(i+\xi),j}\right]^2 \; j = 1,2,3,\ldots,20, \xi < L, \xi \neq 0, \quad (6)$$

where α_j^{ξ} indicates the correlation factor. Finally, a PsePSSM can be expressed from a protein sequence and creates a $20 + 20 \times \xi$ dimensional feature vector.

$$M_{\text{PsePSSM}} = \left[P_1, P_2, \ldots\ldots, P_{20}, \alpha_1^{\xi}, \alpha_2^{\xi}, \ldots, \alpha_{20}^{\xi}\right]^2 \quad (7)$$

This PsePSSM can able to generate a uniform dimensional vector from different lengths of protein sequences in the dataset after extracting features. Here, we set $\xi = 20$ after performing the optimization function for each training set by fivefold CV. Therefore, a $20 + 20 \times 20 = 420$ dimensional feature vector was generated from a protein sequence.

Pseudo AA composition (PseAAC). To consider the order information of protein sequence, Chou [66] introduced a method called PseAAC to extract the features. It can represent AAC information and AA order information both. This method is widely utilized in bioinformatics research, including protein–protein interaction and DTI [67–69] prediction and so on. The following formula can express the features of PseAAC:

$$X = [x_1, x_2, \ldots, x_{19}, x_{20}, x_{20+1} \ldots, x_{20+\lambda}]^T \; (\lambda < L), \quad (8)$$

where L is the length of the given protein sequence. Each of the components is shown as follows:

$$X_v = \begin{cases} \frac{F_\delta}{\sum_{k=1}^{20} F_\delta + W\sum_{k=1}^{\lambda}\Psi_k}, 1 \leq\leq 20 \\ \frac{W\Psi_\delta}{\sum_{k=1}^{20} F_\delta + W\sum_{k=1}^{\lambda}\Psi_k}, 20+1 \leq\leq 20+\lambda \end{cases}, \quad (9)$$

where X represents a feature vector and W indicates the weight factor with a value of 0.05. F_δ represents the frequency at δ-th AA in the protein sequence. We can see from Equation (9) that the first 20 is the frequency of occurrence in protein sequence, and λ is the sequence-related factors that reveal different stages of AA sequence information. It's obtained via the physicochemical properties of AA. In this study, the ranges of the parameter δ are from 0 to 50. Based on the model accuracy of prediction results, the optimal δ parameters can be determined from different parameters setting.

Dipeptide composition. Dipeptide composition (DC) extracts and calculates the frequency of two consecutive AA residues from protein sequence [70]. The sequence encoding technique removes different features involving n-th contiguous AA residues of target protein sequences and computes the occurrences in the sequences. Compared with the AAC, DC considers the coupler effect among adjacent residues; therefore, DC represents not only AA composition information but also the full sequence information of AA. So, DC is one of the best feature extraction techniques and generally generates a 400-dimension feature vector.

$$d = [d_1, d_2 d_3, \ldots, d_{400}]^T, \quad (10)$$

where $d_i (i = 1,2,3,\ldots,400)$ represents residue probability, defined as:

$$d_i = \frac{m_i}{M}, i = 1,2,3,\ldots.,400, \quad (11)$$

where m_i represents residue number and M indicates all possible residue number. For this technique, we used the PyBioMed package [71] to encode protein features where each sequence generates 400 features.

Data balancing technique

As mentioned earlier, our experimental drug–target datasets are highly imbalanced. If such datasets are considered to train the classifier, the model could fail to show accurate prediction performances. Therefore, different data sampling techniques have been utilized in the literature to balance the imbalanced dataset, such as SMOTE [26,72], cluster under sampling [47,73] and random under-sampling [35,74,75]. In this study, we develop a new algorithm based on the concept of random under-sampling technique to overcome the imbalanced problem in the datasets. This algorithm input is imbalance data (minority class samples and majority class samples), and after processing, we will get balanced data as final data. The number of input class samples and features/attributes are diffident for four datasets. Because we constructed three individual drug–target feature groups to evaluate the effect of different features on the model. Therefore, a drug features group is combined with three target feature groups and formed complete drug–target feature groups. The first feature group, namely, MACCS + PsePSSM, achieves 386 features, and other groups, namely, MACCS + PseAAC and MACCS + DC, achieves 206 and 566 features, respectively.

Assume, there are n_1 minority data samples and n_2 majority of data samples in the drug–target datasets. Here, we trained an SVM classifier (multi-kernel or single kernel, decided based on the value of a predefined threshold) to learn the values of features of n_1 minority samples, and after that, we apply the same classifier to attain the features from the n_2 majority of data samples. The threshold value depends on the attributes/features of the datasets. If the number of input features is high, we have fitted the minority samples using a multi-kernel classifier; otherwise, we fitted those with a single kernel classifier. Then Euclidean distance calculates from the predicted and the value of the real features. From the list, we keep these Euclidean distances mapped values by the corresponding majority class samples' indices. Then we arrange the list in descending order using calculated Euclidean distance values. Firstly, we choose n_1 samples from the sorted list. The final data build the combination of n_1 class from the original experimental dataset and n_2 majority class nominated by the proposed method. Therefore, effectively, we select those data samples from the majority class, which are out of the way in terms of Euclidean distance from the predicted values. Said differently, our proposed under-sampling technique generally removes the majority data samples, which are similar to the minority class samples, and retains the majority class samples that are located further from the minority class samples. Hence the decision limit becomes more defined along with the resulting balanced dataset becomes more dividable. However, we need to mention that Algorithm 1 performs effectively when there is no overlap between data points.

Feature selection technique

The dimension of the drug–target features is huge, and the amount of samples is not larger than the size of the features, which is called the feature dimensionality problem [76] and may reason for overfitting results. Therefore, a modified feature selection algorithm, called MoIFS, was developed and implemented based on the theory of incremental feature selection (IFS) technique [77] to obtain optimal features or feature subset for helping in DR. Here, the incremental function expedites the feature selection process when features are enormous and

Algorithm 1: Fast Under Sampling- FastUS

1. $n_1 \leftarrow$ number of minority class samples
2: $n_2 \leftarrow$ number of majority class samples
3: μ ← number of attributes
4: $MA[1 \ldots..n_2] \leftarrow$ Majority class Samples
5: $MI[1 \ldots..n_1] \leftarrow$ Minority class Samples
6: **if** μ > threshold **then** // threshold value depends on the attributes/features
7: *Model* ← *mK*(*MI*[1 . . . ..n_1]) // Multi kernel classifier
8: **else**
9: *Model* ← *sK* (*MI*[1 . . . ..n_1]) // Single kernel classifier
10: **end if**
11: *DA* ← {}
12: **for** each $x \in MA$ **do**
13: $\acute{x} \leftarrow model.predict(x)$
14: $d \leftarrow \|x - \acute{x}\|_2$
15: *index* ← x.*index*
16: *DA* ← *DA* ∪ {*d*, *index*}
17: **end for**
18: *SortedList* ← *sort*(*MA*[1 . . . ..n_1], *DA*)
19: $I \leftarrow L[1 \ldots..mi_1]$
20: $X_1 = \{\}$
21: **for** each *index* ∈ *selectedIndices***do**
22: $X_1 = X_1 \cup MA[index]$
23: **end for**
24: $FinalData = X_1 \cup MI[1 \ldots..n_1]$

can scale up without compromising the quality of drug-target features.

The IFS technique is reformed to have a starting point *K* and sequential performance, reducing cutoff *D*, which is indicated as IFS (*K*, *D*) algorithm. Suppose experimental datasets contain *n* features and *m* samples in a binary classification problem. Based on the association's significance, the drug–target features are represented as binary class labels. Here, these features are calculated by the statistical significance values *P* of the *t*-test [78]. Features are indicated as $f_i, i \in \{1,2,3 \ \ldots \ n\}$, according to their probabilistic rank's values. Our algorithm can successively add elements/values to the feature subset until the accuracy of classification decreases consecutively *D* times. The step by step pseudo-code of Algorithm 2 is shown as follows.

LightGBM classifier

LightGBM is one of the new and powerful algorithms in the ML area. It's a gradient boosting [79] framework Gradient Boosting Decision Tree (GBDT) that uses a decision tree algorithm for learning. If a training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where *x* represents the data samples, and *y* represents the class labels. In GBDT, the *F*(*X*) uses to indicate the estimated function and the optimization function of GBDT minimizes the expected value of some specified loss function*L*(*y*, *F*(*x*)).

$$\hat{F} = \text{argmin} E_{x,y}\left[L\left(y, F(x)\right)\right] \tag{12}$$

To reduce the loss function, the iterative criterion used a line search option in GBDT.

$$F_a(x) = F_{a-1}(x) + \xi_a h_a(x), \tag{13}$$

where $\xi_a = \arg\min \sum_{i=1}^{b} L(y_i, F_{a-1}(x_i) + \xi h_a(x_i))$, *a* represents the iteration number and $h_a(x_i)$ is the base decision tree.

If the experimental datasets are large and enormous features, GBDT algorithms cannot achieve satisfactory accuracy and efficiency. This ensemble algorithm's main cost is to find the best split points during the learning of decision trees. Later, Ke *et al*. [80] introduced an effective gradient boosting algorithm called LightGBM using GOSS (gradient-based one-side sampling) and EFB (exclusive feature bundling). In our approach, LightGBM applies GOSS to control the split through computing variance gain. First, rank the gradient values in descending order of the training samples, and top $a \times 100\%$ data samples of larger gradient values are nominated to get a sample subset A. Then rest of the set A^c containing $(1 - a) \times 100\%$samples with smaller gradients. After that, a subset *B* with size $b \times A^c$ is further randomly sampled. Lastly, we split the samples based on the variance gain over the A ∪ B.

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_r \in B_l} g_i\right)^2}{n_r^j(d)}, \right) \tag{14}$$

where $A_l = \{x_i \in A : x_{ij} \leq d\}, A_r = \{x_i \in A : x_{ij} > d\}, B_l = \{x_i \in B : x_{ij} \leq d\}, B_r = \{x_i \in A : x_{ij} > d\}$, g_i is the negative gradient and $\frac{1-a}{b}$ represents the sum of gradients.

LightGBM has an exclusive function to skip unnecessary calculation for 0 (zero) feature values. Our algorithm can able to optimize the histogram technique to ignore the 0 (zero) feature values. However, we add the optimization in LightGBM as a simple function because this process requires extra computation cost and memory to manage the feature tables in the tree growth procedure. In conclusion, LightGBM is a robust implementation of GBDT with EFB and GOSS to increase model efficiency without losing accuracy. GOSS helps to split the optimal node by computing variance gain, and EFB supports the training process

Algorithm 2: Modified Incremental Feature Selection (MoIFS)

Input: The list of ranked features $\{f_1, f_2, f_3, \ldots\ldots, f_n\}$ and the start position k, The consecutive decreasing cutoff id D

Begin
  F_s=1
  B_f=F_s
  D_t=0
  While D_t<=D;
    if Accuracy(F_s) > Accurancy($F_s \cup \{N_f\}$):
      F_s= $F_s \cup \{N_f\}$
      D_t+=1
    else
      B_f= F_s
      D_t=0
  Return B_f
End

of GBDT by removing zero features. Most importantly, LightGBM is an ensemble-based approach. Based on Equation (13), we can get the LightGBM model $F_m(x)$ by weighted combination scheme.

$$F_m(x) = \sum_{m=1}^{m} \beta_m h_m(x), \tag{15}$$

where m is the maximum number of iteration and $h_m$indicates the base decision tree. The implementation code of LightGBM is available at https://github.com/Microsoft/LightGBM. It follows the histogram-based concept and places continuous values into discrete bins responsible for quicker model training and more effective memory usage. Other boosting algorithms grow trees horizontally, LightGBM grows tree vertically that means it grows tree leaf-wise while other algorithms grow level-wise. It contains complex trees by following leaf-wise split technique instead of a level-wise technique, which is the key reason in attaining the best accuracy. However, it can occasionally lead to model overfitting, which can be avoided by setting the max_depth parameter. Moreover, LightGBM offers over 100 parameters, and it is also supporting optimization in parallel learning to compatibility with large datasets.

Prediction assessment

In this study, the performance of the proposed model is evaluated by a 5-fold CV test to construct an effective prediction framework. Our drug–target datasets were roughly separated into five subsets by 5-fold CV validation test. One set was selected from 5 sets as the test set, and the remaining four were considered as the training set, and this process (cross-validation) was repeated 5 times. After averaging the five validation results, the final results are generated from drug-target datasets. To evaluate the impact of resampling methods on CV results, two types of analyses were performed. In the correct CV, the dataset was first split into k folds, the sampling method (under sampling-FastUS) was applied to the training set constituted of the k – 1-fold, and a reduced training set was obtained. In the incorrect CV, different sampling techniques were first applied to the entire dataset (before CV), and CV was applied to the undersampled data. In this research, we applied the first approach to balance the dataset because applying the balancing method before using the cross-fold validation iterations may lead to biased results. For our feature selection technique (MoIFS), cross-validation is done in each fold as same as the data balancing method (FastUS). In each round of CV, the training and testing samples are changed; therefore, it is needed to find the suitable parameters for MoIFS and classifiers based on new samples. We used the following evaluation metrics to intuitively calculate the performance of the proposed model: accuracy (ACC), sensitivity (SEN), specificity (SPE), Matthews correlation coefficient (MCC) and F1 Score.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{16}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{17}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{18}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN})}} \tag{19}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{20}$$

where TP, FP, TN and FN respectively represent the number of true positives, false positives, true negatives and false negatives. Moreover, we use AUC (area under the curve) to measure the generalization performance of our model. Our model is developed using python language (3.6 version) on Pytorch and scikit-learn library and running on Windows Server PC with system configuration 2.30 GHz Intel Xeon Gold 6140 processor and 128 GB of RAM.

Results and discussion

In this section, we describe the experimental results of our proposed method for detecting new DTIs. We implemented all the techniques, i.e. features extraction, data balancing, classifiers of the proposed model in Python language (Python 3.6 version) using Scikit-learn library, ChemoPy library and LightGBM python package. All the experimental implantations were performed on a high-performance computer with a processor 2.30 GHz Intel Xeon Gold 6140 CPU and 256 GB RAM provided by the Computational Intelligence Lab, UESTC.

Selection of parameter of PseAAC and PsePSSM

Selecting the appropriate parameters of the feature extraction techniques plays a crucial role in constructing a prediction

Table 3. Prediction Performance of DTI for different λ parameter values on the training data set

λ	λ = 1	λ = 2	λ = 3	λ = 4	λ = 5	λ = 6	λ = 7
AUC	0.7345	0.8256	0.8756	0.8745	0.8976	0.8867	0.9045
ACC	0.7102	0.7356	0.7798	0.7865	0.8067	0.7867	0.8105
λ	λ = 8	λ = 9	λ = 10	λ = 11	λ = 12	λ = 13	λ = 14
AUC	0.8999	0.9260	0.9156	0.9104	0.8978	0.9301	0.9381
ACC	0.8065	0.8253	0.8205	0.8035	0.8029	0.8288	0.8439
λ	λ = 15	λ = 16	λ = 17	λ = 18	λ = 19	λ = 20	λ = 21
AUC	0.9345	0.9385	0.9434	0.9325	0.9330	0.9510	0.9428
ACC	0.8368	0.8450	0.8404	0.8299	0.8395	0.9032	0.8478

Table 4. Prediction performances of DTI for different ξ parameter values on the training dataset

ξ	ξ = 1	ξ = 2	ξ = 3	ξ = 4	ξ = 5	ξ = 6	ξ = 7	ξ = 8	ξ = 9	ξ = 10
AUC	0.8845	0.9023	0.9087	0.9067	0.9167	0.9254	0.9180	0.9279	0.9199	0.9289
ACC	0.8615	0.8708	0.8803	0.8788	0.8801	0.8875	0.8840	0.8908	0.8898	0.8876
ξ	ξ = 11	ξ = 12	ξ = 13	ξ = 14	ξ = 15	ξ = 16	ξ = 17	ξ = 18	ξ = 19	ξ = 20
AUC	0.9208	0.9312	0.9330	0.9223	0.9376	0.9289	0.9387	0.9456	0.9489	0.9656
ACC	0.8889	0.8980	0.8976	0.8972	0.8999	0.8982	0.9056	0.9089	0.9096	0.9264

model. When applying the PseAAC and PsePSSM method to extract features from the protein sequence, both selection parameters λ and ξ have a significant effect on the prediction performance of the feature extraction. Here, λ represents the protein sequence information (order information of the amino acid), and ξ represents nothe protein sequence information and contains the protein evolutionary information. Most importantly, if we select the big values of λ and ξ, the protein dimension will be high, which can generate redundant features and affect the prediction task. Conversely, if those parameters' values are too small, the techniques can produce too little protein information, leading to erroneous results.

To find the best value of parameterλ in the prediction model, we set to a range of values is 0 to 21. For different λ values, the LightGBM is applied as a classifier for the prediction task. The 5-fold-CV technique is used to test the model, and the average prediction results are obtained, as listed in Table 3. Moreover, we choose the value of ξ to be 1 to 20 to find the optimal parameters of the prediction model. Similarly, the same classifier and validation technique is used to obtain the prediction performances of the drug-target data, as listed in Table 4. The prediction AUC and ACC value of the model with different changing λ values are shown in Table 3. Whenλ = 17, the AUC and ACC value reached the average 0.9434 and 0.8404, respectively. When λ = 20 compared with λ = 1, the prediction AUC and ACC were improved by 21.65% and 19.30%, respectively. At λ = 10, λ = 11and λ = 12, the prediction AUC and ACC value decreased, but at λ = 13, the AUC and ACC started to increase. Therefore, λ = 20is nominated as the best value of the feature parameter. Because, when λ = 21, the predicted values start to decrease. By PseAAC technique, each protein sequence can get 20 + λ= 20 + 20 = 40 feature vector.

We can see from Table 4, by changing the value of parameter ξ, the AUC and ACC value in the drug-target training data also changes. When the parameterξ = 20, the AUC value reaches a maximum of 0.9656. Among them, comparing ξ = 20 and ξ = 1, AUC and ACC are 8.11% and 6.49% higher, respectively. At ξ = 7 and ξ = 9, AUC and ACC decreased, but the overall performance is rising. Here, the parameter ξ = 20 is better than ξ = 19, and AUC and ACC are increased by 1.67% and 1.68%, respectively. In this study, the best ξ = 20 value is nominated for the model. Each target sequence produces a $20 + 20 \times 20 = 420$ feature vector by using the PsePSSM technique.

The effect of multiple feature extraction

Using perfect mathematical models (feature extraction techniques) to describe and implement it is essential for DTIs research to extract appropriate features from drug chemical structures and protein sequences. In this study, the model is trained by four feature extraction techniques, such as MACCS, PseAAC, PsePSSM and DC. PseAAC is a sequence information-based feature technique that represents information regarding amino acid (AA) sequence order and length of a protein sequence. PsePSSM contains evolutionary information of protein sequence, and DC calculates the frequency of the AA and the frequency of the AA pair. PseAAC, PsePSSM and DC are utilized to generate the protein feature vectors of 40, 220 and 400 dimensions, respectively. The four types of features are provided in the LightGBM classifier for predicting DTIs, and the performance results of feature extraction methods are listed in Table 5. Moreover, we use the receiver-operating characteristic (ROC) curve to represent and compare the performances of the model with various features. Figure 3a–d is the ROC curve generated from the DrugBank datasets under three feature extraction techniques.

We can see from Table 5 that the AUC values of different feature extraction algorithms are also different. For the Enzyme dataset, the model used the MACCS+PsePSSM, MACCS+PseAAC and MACCS+DC features to attain AUC of 0.9656, 0.9510 and 0.9434, respectively. The ACC values of MACCS+PsePSSM, MACCS+PseAAC and MACCS+DC are 0.9264, 0.9032 and 0.8939, respectively. The highest AUC and ACC values were achieved by the MACCS+PsePSSM feature. For the Enzyme dataset, MACCS+PsePSSM gain more comprehensive features from protein sequence, which helps identify DTIs, and the performance results are better than the other two feature extraction techniques. In the case of IC, GPCR and NR datasets, the AUC and ACC values for individual feature groups are also listed in Table 5. Figure 3 shows that the ROC curves of different

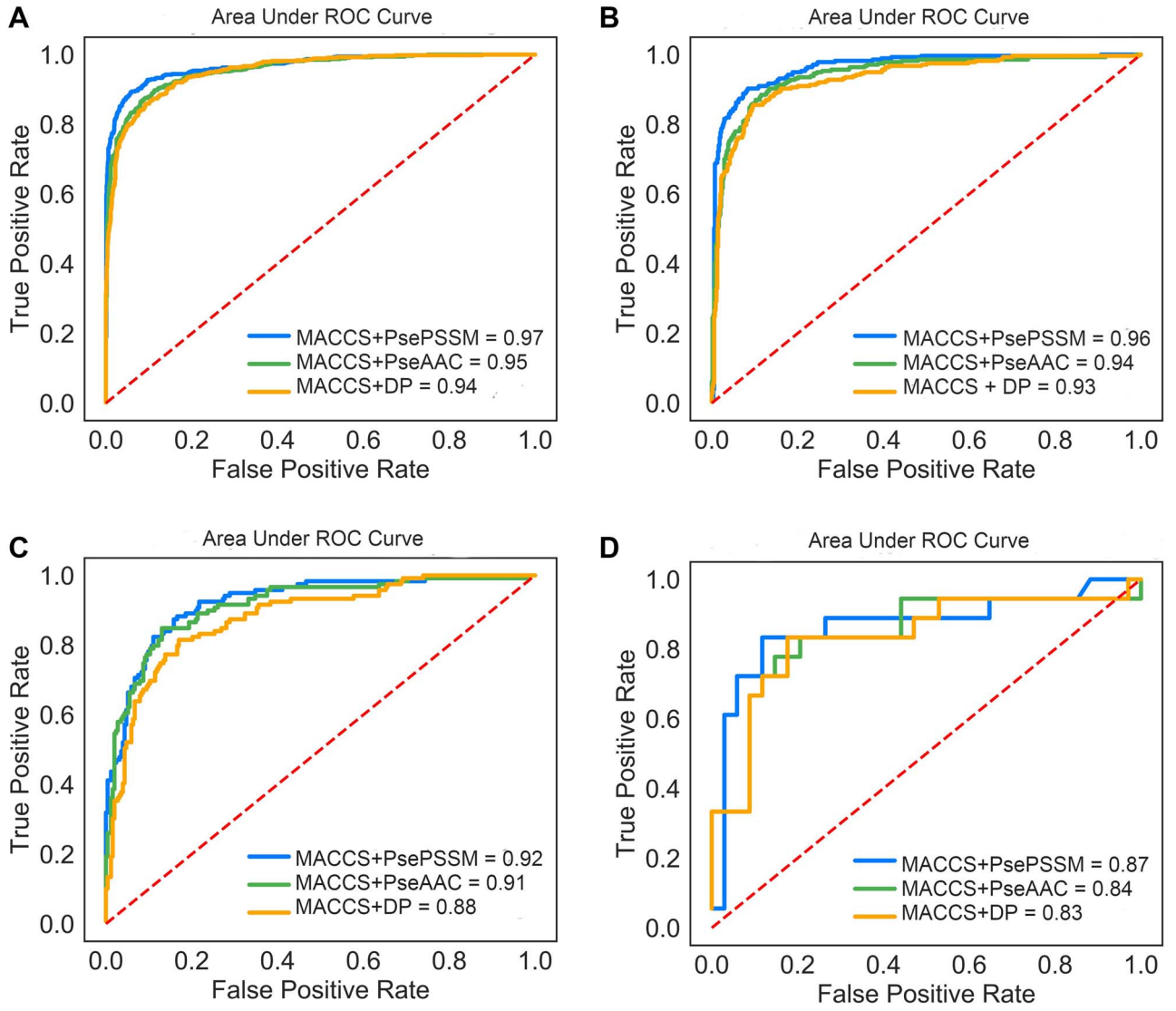

Figure 3. ROC curves of LightGBM classifiers using MACCS+PsePSSM, MACCS+ PseAAC and MACCS+ DC feature group on the datasets: (**a**) EN, (**b**) IC, (**c**) GPCR and (**d**) NR.

Table 5. Performance of different features on benchmark datasets

Datasets	Features	AUC	ACC
EN	MACCS + PsePSSM	0.9656	0.9264
	MACCS + PseAAC	0.9510	0.9032
	MACCS + DC	0.9434	0.8939
IC	MACCS + PsePSSM	0.9612	0.9150
	MACCS + PseAAC	0.9410	0.8859
	MACCS + DC	0.9323	0.8737
GPCR	MACCS + PsePSSM	0.9249	0.8629
	MACCS + PseAAC	0.9134	0.8522
	MACCS + DC	0.8823	0.8306
NR	MACCS + PsePSSM	0.8652	0.8653
	MACCS + PseAAC	0.8445	0.8189
	MACCS + DC	0.8350	0.8076

feature extraction techniques show different coverage areas; the ROC curve using the MACCS+PsePSSM feature covers the largest area, which is significantly higher than the other techniques.

For all the datasets, the MACCS+PsePSSM feature achieved higher prediction performance than MACCS+PseAAC and MACCS+DC. From Table 5, we can see that the features MACCS+PsePSSM, MACCS+PseAAC and MACCS+DC obtained the best performances for EN dataset, followed by IC, GPCR and NR. Most importantly, NR shoed the most unrhymed prediction performance for three feature where MACCS+PsePSSM received a little higher performance than the MACCS+PseAAC and MACCS+DC. Moreover, EN dataset attains the most top results for the DTI features. In contrast, the predictive model achieved the highest AUC of 0.9656 for MACCS+PsePSSM, representing the impact of a protein's evolutionary-based information and structural properties. The MACCS+PsePSSM features also show almost the same performance results for EN and IC. This study's main goal is to compare and examine the effectiveness of different features and determine the most useful feature from the benchmark dataset. Our experiments disclose that MACCS+PsePSSM features more information and shows a significant role in predicting DTIs than MACCS+PseAAC and MACCS+DC feature descriptors.

Table 6. Comparison of prediction results on a balanced and unbalanced dataset

Datasets	Sampling method	Evaluation metrics					
		AUC	ACC	SEN	SPE	MCC	F1
EN	Without resampling	0.9412	0.9067	0.9245	0.8578	0.8289	0.9235
	With FastUS	0.9656	0.9264	0.9456	0.9323	0.8987	0.9347
IC	Without resampling	0.9200	0.8989	0.9056	0.8567	0.7890	0.9056
	With FastUS	0.9612	0.9150	0.9078	0.9234	0.8456	0.9134
GPCR	Without resampling	0.8745	0.8590	0.8765	0.7689	0.6990	0.8800
	With FastUS	0.9249	0.8629	0.8789	0.8765	0.7823	0.8957
NR	Without resampling	0.8353	0.8256	0.9089	0.7289	0.7012	0.8901
	With FastUS	0.8652	0.8653	0.8767	0.8976	0.8234	0.8845

The effectiveness of the data balancing techniques

The imbalanced dataset can be responsible for the biased results. In this study, drug–target dataset is a severe imbalance. The number of known DTI (positive samples) is significantly smaller than that of unknown DTI (negative samples), which is the cause of the reduced performance results of the prediction model. To balance the datasets and improve the ability of the model, we used the FastUS technique as a balancing method with the LightGBM classifier. Here, the model compares the balanced and unbalanced datasets to evaluate the efficiency of the FastUS technique with LightGBM classifier, the experimental results shown in Table 6.

We can see from Table 6 that the model obtains different prediction performances on a balanced (With FastUS) and unbalanced (Without Resampling) dataset. The results show significant advantages on the AUC, ACC, SEN, SPE, MCC and F1 evaluation index after applying the FastUS algorithm. On the unbalanced distribution (Without Resampling), the number of positive instances and negative instances are 2926 and 292 554 (as an example for EN dataset); respectively, the positive instances is less than the negative instances. Using the EN dataset, the MODEL obtained AUC values of 0.9656 for balanced data, 0.9412, for unbalanced data. In the case of the IC dataset, the MODEL achieved AUC values of 0.9612 and 0.9200, for balanced and unbalanced data, respectively. For the GPCR dataset, MODEL yielded an AUC of 0.9249 for balanced, 0.8745 for unbalanced data. Similarly, AUC values of MODEL using NR data are 0.8652for balanced, 0.8353 for unbalanced. In conjunction with balanced and unbalanced data, additional performance metrics, including ACC, SPE, MCC and F1 of the MODEL, are mentioned in Table 6. In the case of EN, the prediction results of ACC, SEN, SPE, MCC and F1 on balanced data are 0.9264, 0.9456, 0.9323, 0.8987 and 0.9347, which are 1.99%, 2.11%, 7.45% 6.98% and 1.12% higher than unbalanced, respectively. It shows that the FastUS technique can obtain a comparatively effective performance. In the case of IC, GPCR and NR datasets, the ACC, SEN, SPE, MCC and F1 results for balanced and unbalanced data are also shown in Table 6. In Figure 4a–d, the ROC curves of data balancing techniques show different coverage areas, the ROC curve using FastUS covers the largest area, which is higher than the without resampling techniques.

It can summarize few clarifications from the above discussion: Firstly, the balanced dataset with FastUS significantly outperforms the unbalanced dataset in the case of ROC curves. Secondly, the performance of the LightGBM classifier has been improved after utilizing the FastUS. More specifically, the results significantly improve for all four datasets on the SPE and MCC metrics, which has worse results for unbalanced data, especially for the NR dataset. Thirdly, FastUS is the effective method for this paper to identify DTIs, since it boosts the prediction ability and reduces the model biasness for drug–target datasets.

Table 7. The prediction results of the MoIFS on EN dataset

Number of Features	AUC	ACC
826	0.9605	0.9098
756	0.9479	0.9077
669	0.9486	0.9178
605	0.9656	0.9264
550	0.9477	0.9056
402	0.9372	0.8934
378	0.9331	0.9056
311	0.9513	0.9023
255	0.9451	0.8967
201	0.9522	0.9034
153	0.9342	0.8812
112	0.9352	0.8867
50	0.9224	0.8745

The influence of feature selection techniques

Feature selection is an essential technique for selecting optimal feature subset in the area of pattern recognition and data processing. It is a combinatorial optimization problem that can increase the prediction ability of the predictive model. Different feature selection techniques have been extensively applied with drug–target datasets in recent studies. Generally the feature selection algorithm assesses feature subsets of drug–target based on the classification algorithm. It provides individual fitness levels and evaluation indicators based on prediction accuracy, to successfully eliminate redundant information in drug-protein features and extract the data for each protein. The experimental results of the drug-target dataset with various feature dimensions are listed in Table 7.

We can see from Table 7; best prediction effect can't get with our dataset when considering full drug–target features. Therefore, it's better to remove some features from the experimental datasets. Moreover, the AUC and ACC scores of the 201 features, 311 features and 826 features are lightly better than the beginning 50 features. Most essential indicators AUC and ACC show advantages result for the 605 features. In the case of AUC, the performance of 605 features is 1.34% high than the 201 features, 1.43% high than the 311features and 0.51% high than the 826 features. In the case of ACC, the ACC of the 605 features are 2.30%, 2.41% and 1.66% higher than the 201,311 and

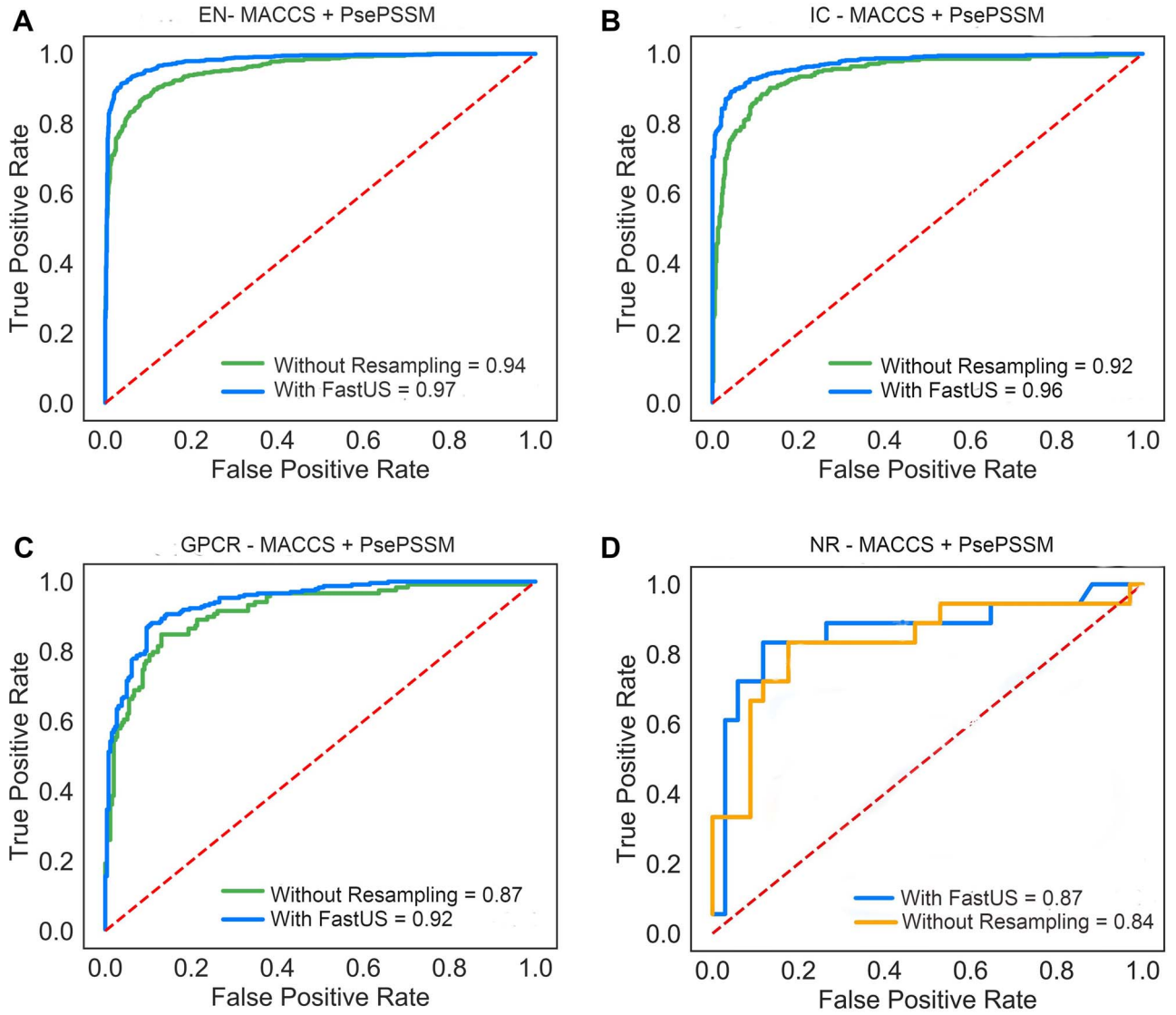

Figure 4. ROC curves of the MACCS+ PsePSSM feature group using without resampling and With FastUS techniques on the datasets: (**a**) EN, (**b**) IC, (**c**) GPCR and (**d**) NR.

826 features. For our experiments, 605 selected features are the optimal dimension by the MoIFS algorithm.

Moreover, to verify the effectiveness of the MoIFS, we compare the prediction performance of MoIFS with other feature selection techniques such as GA [50], PCA [81,82] and Relief [51] under the 826 dimension feature. The prediction result of these three algorithms is shown in Table 8. The best feature dimensions selected by GA, PCA, Relief and MoIFS are 456, 635, 546 and 605, respectively, where AUC values of the prediction model are different. The AUC of 605 for MoIFS is 0.9656, which is 5.55%, 5.22% and 3.41% higher than the GA, PCA and Relief. We can see all the feature selection algorithms can't show the improved results for drug–target datasets, and the MoIFS is a suitable algorithm here. In Figure 5, the ROC curves show the comparison of the robustness of three different feature selection methods where coverage area by the MoIFS is also the largest. Therefore, MoIFS can use this study to reduce the features effectively, improve the prediction performance and reduce the experimental cost.

Selection of model classifier

This study focuses on four classifiers: Random Forest (RF) [83], SVM [84], XGBoost [33] and LightGBM [80]. Here, RF applies 200 trees, and the iterations number of XGBoost is 20. The prediction performances for four classifiers are tested under the cross-validation test, the results of the predictive model shown in Table 9. To make a clear comparison of prediction effects, the results graph of the EN dataset shows in Figure 6. After analyzing the prediction results of the DT dataset from Table 9 that the highest results of AUC, ACC, SEN, SPE, MCC and F1 obtained by the LightGBM algorithm are 0.9656, 0.9264, 0.9456, 0.9323, 0.8987 and 0.9347, respectively. The overall prediction ACC of RF, SVM, XGBoost and LightGBM is 0.8289, 0.8756, 0.9167 and 0.9264, respectively. LightGBM ACC is 0.90%, 5.07% 9.07% higher than that attained by XGBoost, SVM and RF classifiers. The SPE of the LightGBM classifier is 1.89%, 5.79% and 11.89% higher than the XGBoost, SVM and RF classifiers, respectively.

Table 8. The comparison of different feature selection algorithms on EN dataset

Algorithm	Features	Evaluation metrics					
	AUC	ACC	SEN	SPE	MCC	F1	
GA	456	0.9101	0.8765	0.8765	0.8543	0.8587	0.8798
PCA	635	0.9134	0.8827	0.8976	0.8654	0.8478	0.8876
Relief	546	0.9315	0.9024	0.9178	0.8933	0.8756	0.9056
MoIFS	605	0.9656	0.9264	0.9456	0.9323	0.8987	0.9347

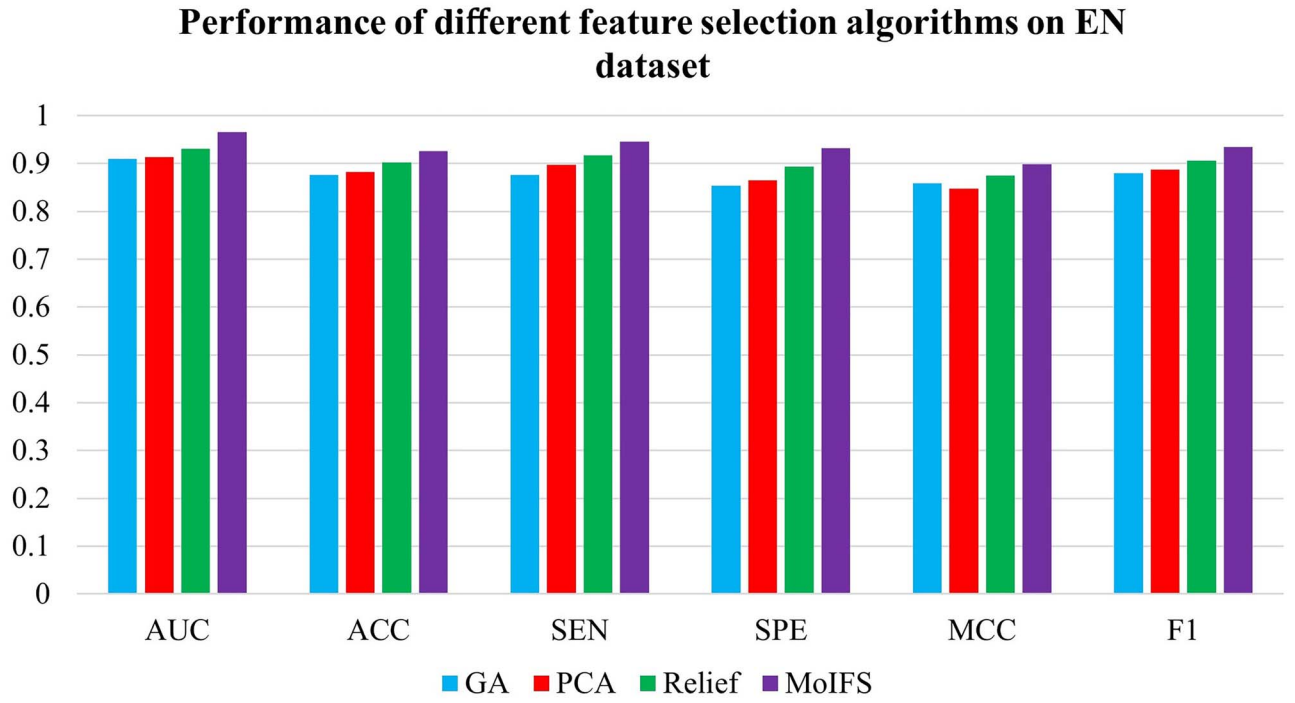

Figure 5. Performance comparison of different feature selection techniques on EN dataset.

Table 9. Performance of different classifiers on EN dataset

Algorithm	Evaluation metrics					
	AUC	ACC	SEN	SPE	MCC	F1
RF	0.8749	0.8289	0.8345	0.8134	0.7999	0.8134
SVM	0.9149	0.8756	0.8834	0.8744	0.8467	0.8751
XGBoost	0.9546	0.9167	0.9312	0.9134	0.8876	0.9199
LightGBM	0.9656	0.9264	0.9456	0.9323	0.8987	0.9347

Figure 6 displays the bar graph of the EN data for four classifiers. We can see from Figure 6 that the prediction values found by the LightGBM classifier have the best AUC score of 0.9656, which is 1.10%, 5.07% and 8.749% higher than that of the XGBoost, SVM and RF classifiers, respectively. Through comparing the bar graph, AUC, ACC, SEN, SPE, MCC and F1 values of four classifiers on the drug–target data, robust performances with generalization ability is considered for a classifier to construct our model. Here, the RF classifier is practical and straightforward, but the results generated by this classifier are relatively weak because of the larger calculation amount. SVM and XGBoost are tree-based classifiers, but there are still some challenges in the prediction task. The prediction performance of the LightGBM classifier is superior to the other three classifiers. Therefore, we select the LightGBM classifier as a classification algorithm for this study.

Comparison with other methods

As mentioned earlier, different feature extraction, feature selection, data balancing and classifiers have been used for detecting interactions between a drug and target protein. To compare the effectiveness of our method, we consider five drug–target methods under the AUC values for the same datasets. Here, those existing methods also utilized 5-fold CV as a key performance metric for four datasets. Our model has compared with that of Huang *et al.* [85], Mousavian *et al.* [35], Li *et al.* [41], Wang *et al.* [59] and Meng *et al.* [86] under the AUC values. They

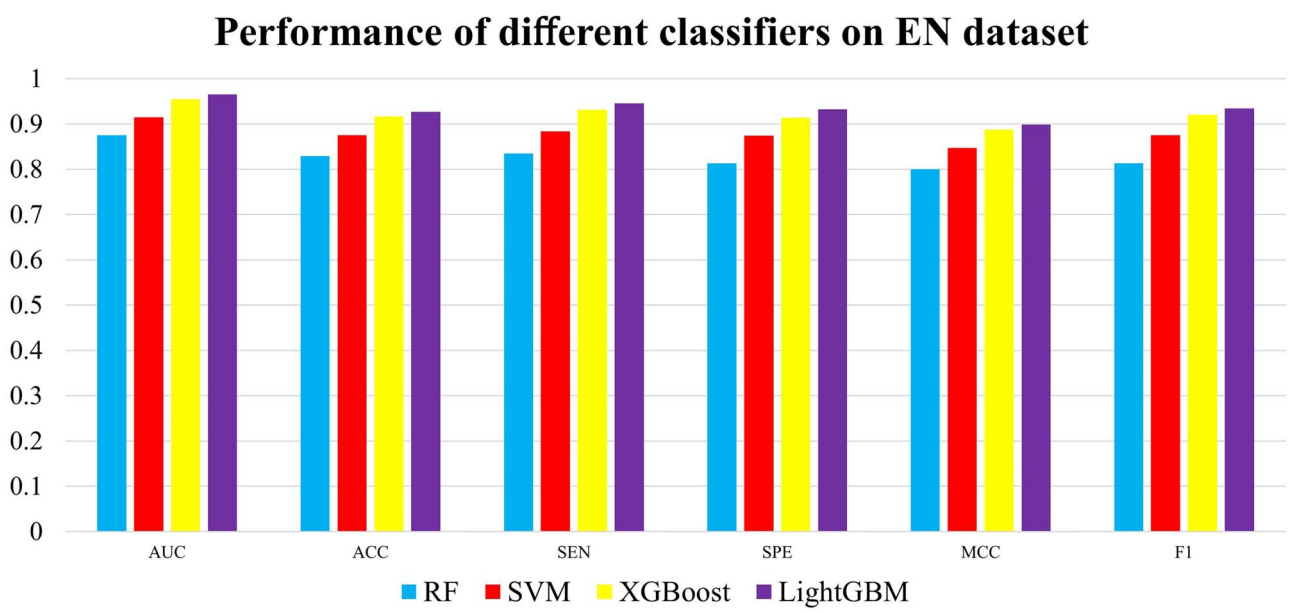

Figure 6. Performance comparison of different classifiers on EN dataset.

Table 10. Comparison of MODEL with existing methods on four datasets

Dataset	Huang *et al.* [85]	Mousavian *et al.* [35]	Li *et al.* [41]	Wang *et al.* [59]	Meng *et al.* [86]	Our method
EN	0.9040	0.9480	0.9288	0.9425	0.9773	0.9656
IC	0.8510	0.8890	0.9171	0.9107	0.9312	0.9612
GPCR	0.8990	0.8720	0.8856	0.8743	0.8677	0.9249
NR	0.8430	0.8690	0.9300	0.8176	0.8778	0.8652

made the negative samples with the same size as the positive sample datasets using random sampling techniques to solve the imbalance problem. The AUC values of the five methods with our proposed method are reported in Table 10, where it can be seen that the AUC values of PreDTIs are significantly higher than the six exiting methods.

Here, Huang *et al.* [85] used the extremely randomized tree (ER-Tree) algorithm as a classifier where Pseudo-SMR and fingerprint descriptor are applied to represent the biological evolutionary information and molecular substructure information of drug-target. The results show that this method achieves a prediction AUC of 0.9040, 0.8510, 0.8990 and 0.8430, respectively. To investigate the influence of a negative selection scheme on the prediction performance, Mousavian *et al.* [35] developed an SVM-based model with Bigram-PSSM model and Fingerprints methods for predicting DTIs. This method's average AUC values on four datasets are 0.9480, 0.8890, 0.8720 and 0.8690, respectively. Li *et al.* [41] introduced a DVM classifier using highly discriminative information of DTIs. The evolutionary information is retained from PSSM, and then the LBP technique is utilized to compute the LBP histogram descriptor. Wang *et al.* [59] proposed an ML framework for identifying DTIs from drug chemical information and protein sequence using a deep-stacked autoencoder, which can effectively generate raw information. This proposed framework shows some advantages that it can attain the hidden data from target sequences and extract representative features through multiple layers iterations. The prediction results of AUC on data are 0.9425, 0.9107, 0.8743 and 0.8176. Meng *et al.* [86] introduced a method, namely PDTPS to detect DTIs. This approach combines PSSM, BIGP and PCA with RVM. This approach achieved an average accuracy of 97.73%, 93.12%, 86.78% and 87.78% on EN, IC, GPCR and NR, respectively.

Average AUC values of PreDTIs on EN, IC, GPCR and NR are 0.9656, 0.9612, 0.9249 and 0.8652, respectively. We see that PreDTIs significantly outperformed the exiting approaches in terms of the AUC metric. However, Meng *et al.* [86] achieved little better results than us for EN and NR datasets. Our method's high prediction performance effectively features extraction techniques that extract more discriminative features for molecules and proteins. Furthermore, our balancing method perfectly manages the imbalance problem in the datasets, and feature selection techniques reduce the unwanted features, which also the main reason for better performance by the LightGBM classifier, indicating better performance for identifying the new DTIs.

New DTIs prediction

After investigating our predicted drug-target pairs, the PreDTIs model only does not attain the high probability scores in terms of the AUC metric but also predicted interactions are biologically realistic. It is important to remember that the interacted datasets used in this study were collected from few years old databases. Most importantly, those interacted datasets still exist and unchanged; therefore, we can verify our newly predicted drug-target pair with an updated version of the database. Meanwhile, many new interactions have been discovered by the wet-lab experiments and stored those interactions in the updated version of databases such as DrugBank [11], ChEMBL [10] and KEGG [87].

The model is trained using four benchmark datasets, and the non-interacting pairs are labeled based on their prediction

probability. After training, our model could predict new interactions from non-interacting samples. We selected only the top new pairs that achieved the highest prediction probability and listed in Supplementary Tables 1–4. The predicted pairs, which are achieved at least 82% prediction probability by our model, can be considered correct predictions and found in the existing databases. Here, the current version of KEGG and DrugBank as of 7 June 2020 is considered to search and verify our interactions.

In particular, we investigated one target protein, Prothrombin (P00734). Figure 7 shows the top predicted for Prothrombin. In our experiments, Prothrombin has a total of 30 predicting drugs with our method, and seven were effectively identified as presented in Figure 7. The unconfirmed predictions might be correct after investigating their possibility. These experimental results disclose that our PreDTIs approach could also be utilized to predict the new drugs for severe acute respiratory syndrome 2 (SARS-CoV-2), indicating the proposed techniques are more practical in real applications.

Discover new drugs to treat COVID-19 patients

Coronavirus disease is a critical health challenge across the world since December 2019. Until now, 14 450 472 cases, with 605 587 confirmed deaths, have been reported in 215 countries. The confirmed COVID-19 cases are quickly spreading, and the incidence rate is growing worldwide. Peoples affect by this virus, and there is no vaccine in the globe yet for treatment against this virus. Therefore, there is a serious need to find effective anti-COVID-19 agents for the prevention of the epidemic and control viral infections. Drug repurposing of antiviral molecules is a strategy employed in COVID-19 treatment, representing a valid alternative of the vaccine in this short period. Most of the existing drugs are used for repurposing in the treatment of the COVID-19, and the medical community knows their therapy and toxicity information in humans. Recently, several drugs are selected, which shows promising inhibitors against COVID 19 protease. There are some molecular docking (computational methods)-based methods that have been applied to FDA-approved drugs or antiviral drugs to identify useful molecules on viral proteins of SARS-CoV-2. But ML [88] based method can be a helpful tool. In this section, we propose to use our PreDTIs (ML-based DTI) model to identify the most potent drugs against SARS-CoV-2. A step by step of DTI-COVID-19 framework is illustrated in Figure 8, and the details information about this method described as follows:

- Data collection and analysis of SARS-CoV-2: Specimens will be collected from humans. Upper respiratory specimens and lower respiratory specimens need to collect as a respiratory material. If required, other clinical specimens such as blood and stool also can be collected from the patients. During this process, the medical staff should well be trained with enough Standard Operating Procedures for collecting an appropriate specimen from the human body. After that, these collected samples should reach the testing laboratory as early as possible and can be shipped and stored at 2–8°C. The virus analyzer machine will process the specimens and gene expression analysis done to get virus protein sequences.
- Drug and protein datasets collection: FDA-approved drugs or antiviral drugs can be selected as experimental drugs from the DrugBank database, which contains detailed information of drugs compound, and the protein of SARS-CoV-2 can be collected from the NCBI database. There is a specific link (https://www.ncbi.nlm.nih.gov/sars-cov-2/) for SARS-CoV-2 where protein sequence can be collected for experiments purposes.
- Features generation: There are various types of techniques to extract protein sequences and drug chemical structures to represent them into a numerical form such as PubChem fingerprint, MACCS fingerprint, PsePSSM, PseAAC, PSSM-Bigram, etc. The model can choose any of the techniques for the processing of its structure and sequence.
- Applied diffident techniques on datasets: If the experimental datasets are imbalanced ith huge features, then data balancing and feature selection can be applied on the datasets to handle the balance issue and reduce the drug-protein features. Finally, the ML classifier could be applied to the dataset for prediction propose.
- Find effective drug–protein pairs: Different statistical and ML techniques such as Pearson's correlation coefficient, Spearman's correlation coefficient and prediction probability can be used to determine the active pairs from the datasets.
- Suggest drugs for COVID-19: Based on the correlation coefficient and prediction probability scores, the model can show the best drugs against COVID-19.

Discussion

Our proposed method has effectively ranked drug-target pairs and suggests new drugs for exiting proteins. This research's main contribution is to develop a balancing algorithm and feature selection algorithm to handle the class-imbalanced and manage the high dimensionality of data. Moreover, boosting help in adding diversity for the multi-features and improves the prediction performance. The experimental results show that our method outperforms the existing methods. Most importantly, we can clearly see that the model's performance is degraded when the balancing technique is not performed. Besides, there is a significant gain in performance when the proposed DR algorithm is applied in the dataset. Our proposed method is benefitted more by MoIFS in comparison to GA, PCA and Relief. But, the running time of MoIFS is quite high on the EN dataset, our proposed approach's main complexity.

Further, to check the method's ability, we investigated one target protein, prothrombin (P00734), and found a total of 30 (top 7 is shown in Figure 7) predicting drugs. These results indicate that our PreDTIs approach could also be utilized to predict the new drugs for SARS-CoV-2, indicating the proposed techniques are more practical in real applications. Further, new and missing DTIs were identified using our framework. To verify the model efficiency, some known (positive) DTIs were removed from the benchmark data, and their DTIs were recalculated to verify the model accuracy. The results disclose that similar structures of compounds tend to interact with similar types of targets.

Conclusions

With the massive expansion of the era of big data, protein sequence and drug chemical structure data are increasing rapidly in different biological databases. Therefore, ML has become an effective strategy to predict the DTIs. In this article, we present a new ML-based method PreDTIs to predict DTIs. This model is developed by fusing different types of drug and target features with the LightGBM classifier. The key challenges of DTIs prediction include: (1) fully extracting the critical features of drug and protein; (2) the problem of imbalance data

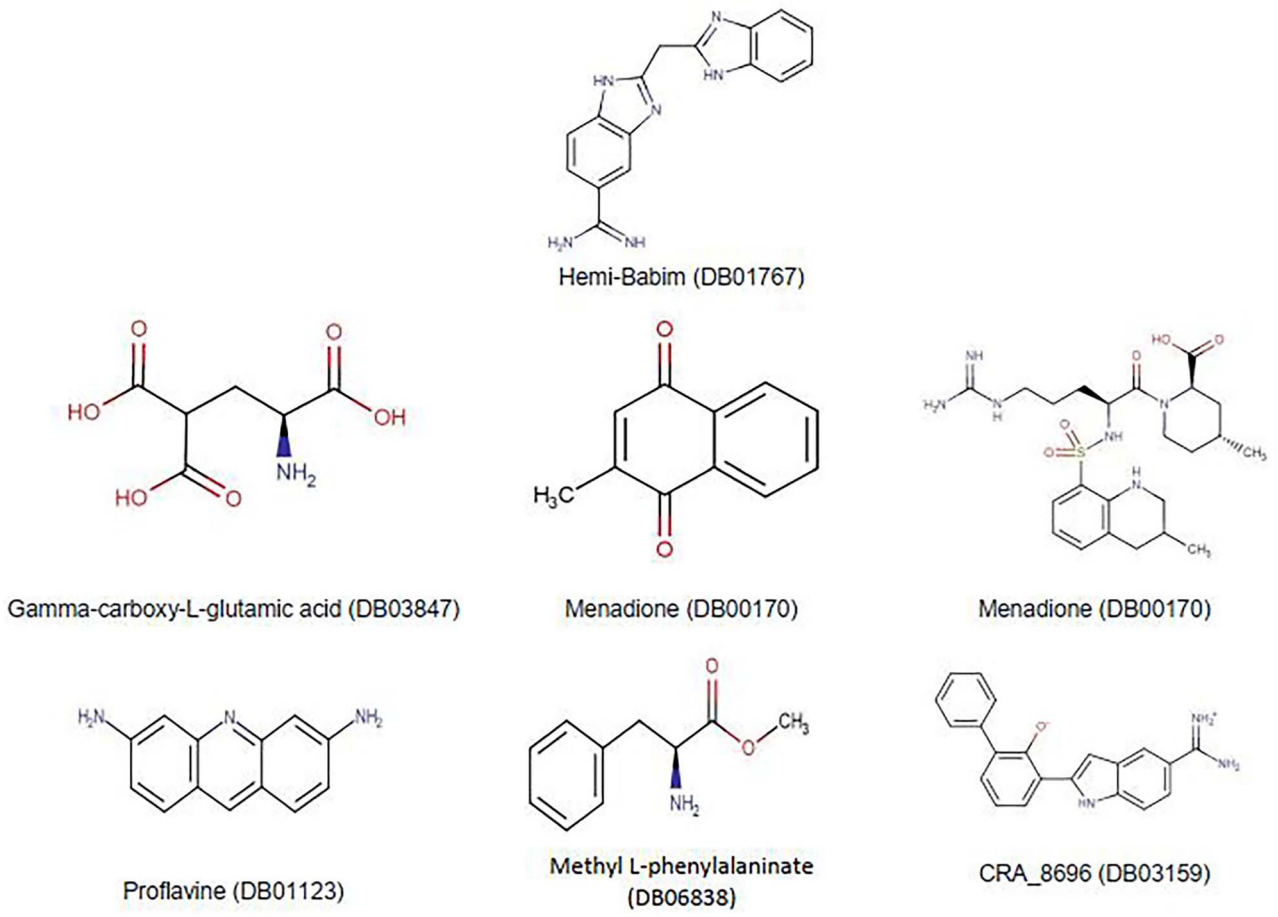

Figure 7. Chemical structures of drugs for prothrombin.

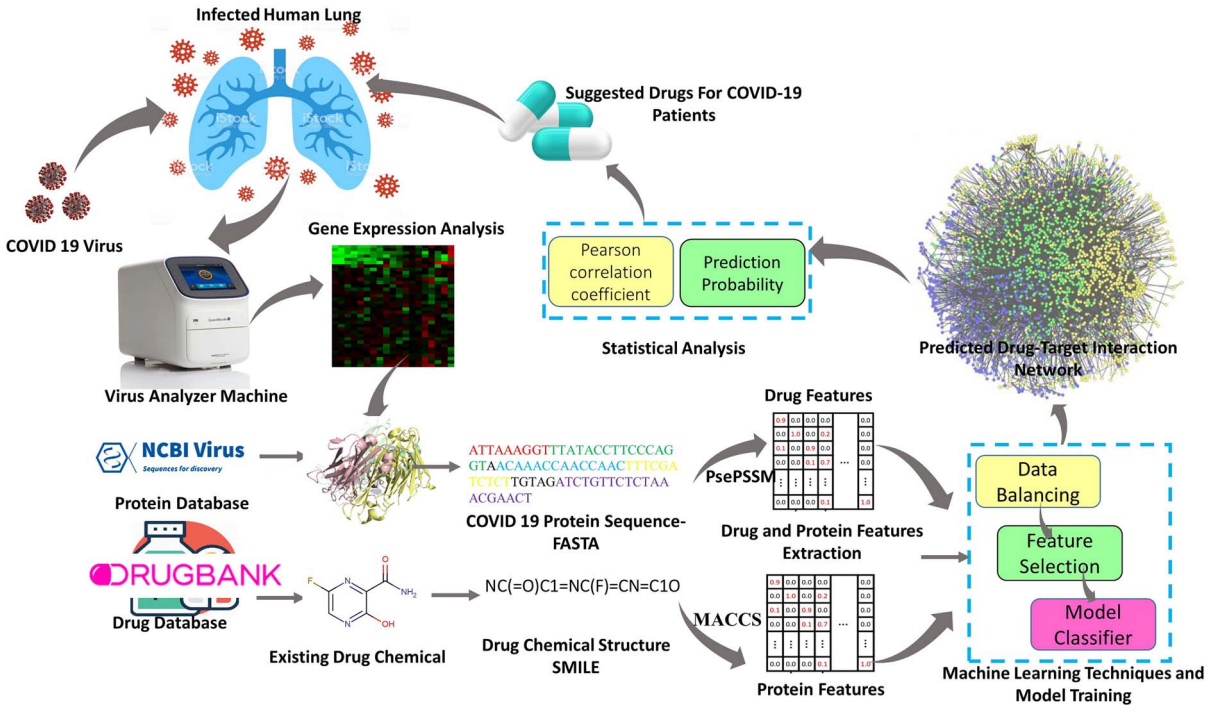

Figure 8. A ML-based method (schematic diagram) to discover new drugs to treat COVID-19 patients.

and (3) feature dimensionality in the dataset. First, PsePSSM, PseAAC, DC and MACCS fingerprint are employed to extract the evolutionary, sequence, and structural information of the target and drug features to identify DTIs. The imbalance data is a common problem in the biological datasets; therefore, we proposed a balancing algorithm based on the concept of random under-sampling techniques, which helps in handling the data-imbalance issue and minimizes the prediction biasness. Most of the related exiting works didn't consider the imbalanced data problem effectively while developing a prediction model. Moreover, we developed a feature selection algorithm to reduce the dimension of the drug–target features to retain the optimal set of features effectively and make the prediction process flexible. This algorithm shows a perfect mechanism to enhance prediction performances for predicting DTIs. Finally, balanced data and optimal features are provided into the LightGBM Classifier to predict unknown DTIs. The 5-fold CV validation is used to evaluate the predictive performance of the proposed method. The prediction results show that the proposed method has improved performance compared with other related existing methods using the same dataset. We believe our method not only useful for the prediction of DTIs but also a robust application in a relevant area such as molecular biology, bioinformatics and proteomics. Here, too, we show how to use our PreDTIs (ML approach) model to identify the most potent drugs against SARS-CoV-2. In the future, we have a plan to consider SARS-CoV-2 datasets for experiments and use a deep learning-based algorithm as a classifier to speed up the model performance in discovering new drugs for COVID-19. The current version of our proposed method is suitable for both the gold standard and DrugBank dataset, but not for heterogeneous data sources. There is a chance to face some core challenges for our model with heterogeneous data because (i) there will be missing information as not all kinds of data will be available for all the targets and drugs and their interactions; (ii) the heterogeneous data sources are not all of the same quality, and data volume can be relatively huge.

Fortunately, with the accumulation of a large amount of health data and the development of ML methods, our algorithm can predict drug-disease treatment associations to minimize the effort and cost. A drug that can interact with multiple target proteins is a pretty common scenario; therefore, it is important to consider the drug repositioning concept to identify treatments for novel diseases. We can provide an example of the drug Gleevec (STI-571 or Imatinib Mesylate), which was initially bound with the fusion gene BCR-ABL associated with blood cancer disease leukemia; later, Gleevec was also bound with KIT and Platelet-derived growth factor (PDGF), it has led to a revolution in the therapeutic approach to gastrointestinal stromal tumors. Many existing drugs may expect to bind with unknown targets for treating novel diseases using our proposed algorithm.

Symbols/notations

Symbol/Notation	Description
MA	Majority Samples in the datasets
MI	Minority Samples in the datasets
μ	Number of attributes
sK	A single kernel is trained
mK	Multiple kernels are trained
ξ	Prediction by model
DA	Distance samples array
L	A Sort list to store the majority samples and distance samples
I	Selected indices from the minority samples
F	A final data output frame
F_s	Feature subset
F_b	Best features from imputed features
D_t	Decrease Times
N_f	Next feature of all features

Key Points

- A new computational model predicts unknown DTIs using protein sequences and drug chemical structures to suggest new drugs for known targets and find new targets for current drugs.
- In the feature extraction stage, generated drug-target features can represent us to their discriminatory nature for patterns related to evolutionary, sequence and structural information that helps predict new DTIs interactions even more effective.
- Imbalanced datasets can lead to losing the model's ability to give accurate decisions where the prediction generally occurs based on the majority class and completely omits the minority class. The proposed FastUS algorithm solves the class imbalance problem in the drug-target datasets.
- High dimensionality features may cause model overfitting. The proposed algorithm MoIFS can help obtain the optimal features based on the theory of IFS where features are enormous and able to scale up without compromising the quality of drug–target features.
- The model achieves the best prediction performance and can suggest potential DTIs, even the effective drug candidates against COVID-19.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

The National Natural Science Foundation of China—Research on New Technology of Core Algorithm (under Grant 61772115).

References

1. He Z, Zhang J, Shi XH, *et al.* Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One* 2010;**5**. doi: 10.1371/journal.pone.0009603.
2. Knowles J, Gromo G. Target selection in drug discovery. *Nat Rev Drug Discov* 2003;**2**:3–9. doi: 10.1038/nrd986.
3. Chen T, Chu Y, Shan X, *et al.* DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Brief Bioinform* 2020;1–15. doi: 10.1093/bib/bbaa205.
4. van de Waterbeemd, Gifford E. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* 2003;**2**:192–204. doi: 10.1038/nrd1032.
5. Johnson DE, Wolfgang GHI. Predicting human safety : screening and computational approaches. *Drug Discov Today* 2000;**5**:445–54.
6. Fakhraei S, Huang B, Raschid L, *et al.* Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**:775–87.

7. Hopkins AL. Predicting promiscuity. *Nature* 2009;**462**:167–8. doi: 10.1038/462167a.
8. Chen X, Yan CC, Zhang X, *et al*. Drug – target interaction prediction : databases, web servers and computational models. *Brief Bioinform* 2016;**17**:696–712. doi: 10.1093/bib/bbv066.
9. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14. doi: 10.1093/nar/gkr988.
10. A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M.\ Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, J.P. Nowotka, M.; Papadatos, G.; Santos, R.; Overington, the ChEMBL bioactivity database: an update, *Nucleic Acids Res* **42** (2013) D1083–90. doi:10.1093/nar/gkt1031.
11. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res* 2011;**39**:D1035–41. doi: 10.1093/nar/gkq1126.
12. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res* 2002;**30**:412–5.
13. Zhu F, Han B, Kumar P, *et al*. Update of TTD: therapeutic target database. *Nucleic Acids Res* 2010;**38**:D787–D791.
14. Szklarczyk D, Santos A, Von Mering, *et al*. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**:D380–4. doi: 10.1093/nar/gkv1277.
15. Jin G, Wong STC. Toward better drug repositioning : prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today* 2014;**19**:637–44. doi: 10.1016/j.drudis.2013.11.005.
16. Bagherian M, Sabeti E, Wang K, *et al*. Machine learning approaches and databases for prediction of drug – target interaction : a survey paper. *Brief Bioinform* 2019;**00**:1–23. doi: 10.1093/bib/bbz157.
17. Keiser MJ, Roth BL, Armbruster BN, *et al*. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**:197–206. doi: 10.1038/nbt1284.
18. Regad L, Reyne C, Spe O. Insights into an original pocket-ligand pair classification : a promising tool for ligand profile prediction. *PLoS One* 2013;**8**. doi: 10.1371/journal.pone.0063730.
19. Cheng AC, Coleman RG, Smyth KT, *et al*. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**:71–5. doi: 10.1038/nbt1273.
20. Combs SA, Deluca SL, Deluca SH, *et al*. Small-molecule ligand docking into comparative models with Rosetta. *Nat Protoc* 2013;**8**:1277–98. doi: 10.1038/nprot.2013.074.
21. Zhu S, Okuno Y, Tsujimoto G, *et al*. A probabilistic model for mining implicit 'chemical compound – gene' relations from literature. *Bioinformatics* 2005;**21**:245–51. doi: 10.1093/bioinformatics/bti1141.
22. Mousavian Z, Masoudi-Nejad A. Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin Drug Metab Toxicol* 2014;**10**:1273–87. doi: 10.1517/17425255.2014.950222.
23. Wu Z, Cheng F, Li J, *et al*. SDTNBI: An integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. *Brief Bioinform* 2017;**18**:333–47. doi: 10.1093/bib/bbw012.
24. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery : methods, tools and databases. *Brief Bioinform* 2018;1–35. doi: 10.1093/bib/bby061.
25. Yu H, Chen J, Xu X, *et al*. A systematic prediction of multiple drug-target interactions from chemical, genomic and pharmacological data. *PLoS One* 2012;**7**. doi: 10.1371/journal.pone.0037608.
26. Mahmud SMH, Chen W, Jahan H, *et al*. iDTi-CSsmoteB : identification of drug–target interaction based on drug chemical structure and protein sequence using XGBoost with oversampling technique SMOTE. *IEEE Access* 2019;**7**:48699–714. doi: 10.1109/ACCESS.2019.2910277.
27. Gönen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;**28**:2304–10. doi: 10.1093/bioinformatics/bts360.
28. Zheng X, Hao Ding HMSZ. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions categories and subject descriptors, in: 19th ACM SIGKDD Int. *International Conference on Knowledge Discovery and Data Mining*. Chicago, IL, USA n.d.1025–33.11–14 August 2013
29. Ezzat A, Zhao P, Wu M, *et al*. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinform* 2016. doi: 10.1109/TCBB.2016.2530062.
30. Bagherian M, Kim RB, Jiang C, *et al*. Coupled matrix – matrix and coupled tensor – matrix completion methods for predicting drug – target interactions. *Brief Bioinform* 2020;**00**:1–11. doi: 10.1093/bib/bbaa025.
31. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst* 2012;**8**:1970–8. doi: 10.1039/c2mb00002d.
32. Lan K, Wang D, Fong S, *et al*. A survey of data mining and deep learning in bioinformatics. *J Med Syst* 2018;**42**:139.
33. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): pp. 785–94. doi:10.1145/2939672.2939785.
34. Hu S, Xia D, Su B, *et al*. A convolutional neural network system to discriminate drug-target interactions*IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2019. doi: 10.1109/TCBB.2019.2940187.
35. Mousavian Z, Khakabimamaghani S, Kavousi K, *et al*. Drug-target interaction prediction from PSSM based evolutionary information. *J Pharmacol Toxicol Methods* 2016;**78**:42–51. doi: 10.1016/j.vascn.2015.11.002.
36. Xiao X, Min JL, Wang P, *et al*. ICDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J Theor Biol* 2013;**337**:71–9. doi: 10.1016/j.jtbi.2013.08.013.
37. G. G. Kiruba B and D.P.Acharjya., Behavioural intention of customers towards smartwatches in an ambient environment using soft computing: An integrated SEM-PLS and fuzzy rough set approach, *Int J Ambient Comput Intell* **11** (2020) 80–111. doi:10.4018/IJACI.2020040105.
38. Yamanishi Y, Araki M, Gutteridge A, *et al*. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**:232–40. doi: 10.1093/bioinformatics/btn162.
39. Yamanishi Y, Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**:246–54. doi: 10.1093/bioinformatics/btq176.
40. Hao M, Wang Y, Bryant SH. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. *Anal Chim Acta* 2016;**909**:41–50. doi: 10.1016/j.aca.2016.01.014.

41. Li Z, Han P, You ZH, *et al*. In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci Rep* 2017;**7**:1–13. doi: 10.1038/s41598-017-10724-0.
42. Rayhan F, Ahmed S, Shatabda S, *et al*. IDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. *Sci Rep* 2017;**7**:1–18. doi: 10.1038/s41598-017-18025-2.
43. Wang L, You Z-H, Chen X, *et al*. RFDT: a rotation Forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. *Curr Protein Pept Sci* 2018;**19**:445–54. doi: 10.2174/1389203718666161114111656.
44. You J, Mcleod RD, Hu P. Predicting drug-target interaction network using deep learning model. *Comput Biol Chem* 2019;**80**:90–101. doi: 10.1016/j.compbiolchem.2019.03.016.
45. Shi H, Liu S, Chen J, *et al*. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 2018;1–14. doi: 10.1016/j.ygeno.2018.12.007.
46. Zhang J, Zhu M, Chen P, *et al*. DrugRPE : random projection ensemble approach to drug-target interaction prediction. *Neurocomputing* 2017;**228**:256–62. doi: 10.1016/j.neucom.2016.10.039.
47. Mahmud SMH, Chen W, Meng H, *et al*. Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Anal Biochem* 2020;**589**. doi: 10.1016/j.ab.2019.113507.
48. FISHER RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936;**7**:179–188.
49. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1993;**24**:417–41.
50. Holland J. Genetic algorithms. *Sci Am* 1992;**267**:66–73.
51. Robnik-Šikonja I. M., Kononenko, theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* 2003;**53**:23–69.
52. Thafar MA, Olayan RS, Ashoor H, *et al*. DTiGEMS +: drug – target interaction prediction using graph embedding , graph mining, and similarity - based techniques, J. *Chem* 2020;**12**:1–17. doi: 10.1186/s13321-020-00447-2.
53. Manoochehri HE, Nourani M. Drug-target interaction prediction using semi-bipartite graph model and deep learning. *BMC Bioinformatics* 2020;**21**:1–16. doi: 10.1186/s12859-020-3518-6.
54. Chen Z, You Z, Guo Z, *et al*. Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model, front. *Bioeng Biotechnol* 2020;**8**:1–9. doi: 10.3389/fbioe.2020.00338.
55. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0 : a major update to the DrugBank database for 2018. *Nucleic Acids Res*46 2018D1074–D1082. doi: 10.1093/nar/gkx1037.
56. Günther S, Kuhn M, Dunkel M, *et al*. SuperTarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;**36**:919–22. doi: 10.1093/nar/gkm862.
57. Kanehisa M, Araki M, Goto S, *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2008;**36**:480–4. doi: 10.1093/nar/gkm882.
58. Schomburg I. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;**32**:431D–3. doi: 10.1093/nar/gkh081.
59. Wang LEI, You Z, Chen X, *et al*. Based method for predicting drug–target interactions by using stacked autoencoder deep neural. *Network* 2017;**24**:1–13. doi: 10.1089/cmb.2017.0135.
60. Ding Y, Tang J, Guo F. Identification of drug-target interactions via multiple information integration. *Inf Sci (Ny)* 2017;**418–419**:546–60. doi: 10.1016/j.ins.2017.08.045.
61. Cao DS, Liu S, Xu QS, *et al*. Large-scale prediction of drug-target interactions using protein sequences and drug topological structures. *Anal Chim Acta* 2012;**752**:1–10. doi: 10.1016/j.aca.2012.09.021.
62. Cao D, Xu Q, Hu Q, *et al*. ChemoPy : freely available python package for computational biology and chemoinformatics. *Bioinformatics* 2013;**29**:1092–4. doi: 10.1093/bioinformatics/btt105.
63. Shen H, Chou K. Nuc-PLoc : a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng Des Sel* 2013. doi: 10.1093/protein/gzm057.
64. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**(2):195–202.
65. Altschul SF, Madden TL, Schaffer AA, *et al*. PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402. http://us.expasy.org/sprot.
66. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;**21**:10–9. doi: 10.1093/bioinformatics/bth466.
67. Jia J, Liu Z, Xiao X, *et al*. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. *J Biomol Struct Dyn*ISSN. 1102 (2015) . doi: 10.1080/07391102.2015.1095116.
68. Zhai J, Cao T, An J, *et al*. Highly accurate prediction of protein self-interactions by incorporating the average block and PSSM information into the general PseAAC. *J Theor Biol* 2017;**432**:80–6. doi: 10.1016/j.jtbi.2017.08.009.
69. Zhu P, Li W, Zhong Z, *et al*. Molecular BioSystems predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Mol Biosyst* 2015;**11**:558–63. doi: 10.1039/C4MB00645C.
70. Khan A, Majid A, Hayat M. CE-PLoc : An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. *Comput Biol Chem* 2011;**35**:218–29. doi: 10.1016/j.compbiolchem.2011.05.003.
71. J. Dong, Z.J. Yao, L. Zhang, F. Luo, Q. Lin, A.P. Lu, A.F. Chen, D.S. Cao, PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions, *J Chem* **10** (2018) 1–11. doi:10.1186/s13321-018-0270-2.
72. Shi H, Liu S, Chen J, *et al*. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 2018;1–14. doi: 10.1016/j.ygeno.2018.12.007.
73. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst Appl* 2009;**36**:5718–27. doi: 10.1016/j.eswa.2008.06.108.
74. Li J, *et al*. Rare event prediction using similarity majority under-sampling technique. *Soft Comput Data Sci* 2017. doi: 10.1007/978-981-10-7242-0_3.
75. Arefeen A, Nimi ST, Rahman MS, *et al*. Neural network-based undersampling techniques, *IEEE Transactions on Systems, Man, and Cybernetics*. 2020;1–10. doi: 10.1109/TSMC.2020.3016283.

76. Chowdhury SY, Shatabda S, Dehzangi A. iDNAProt-ES : identification of DNA-binding proteins using evolutionary and structural features. *Sci Rep* 2017;1–14. doi: 10.1038/s41598-017-14945-1.
77. Liu H, Setiono R. Incremental feature selection. *Appl Intell* 1998;**9**:217–30.
78. Ye Y, Zhang R, Zheng W, *et al*. RIFS: a randomly restarted incremental feature selection algorithm. *Sci Rep* 2017;1–11. doi: 10.1038/s41598-017-13259-6.
79. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* n.d.;**29**:1189–232. doi: 10.2307/2699986.
80. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: *31st Conference on Neural Information Processing Systems (NIPS)* (2017): pp. 3146–54.
81. Karl Pearson FRS. On lines and planes of closest fit to systems of points in space. *Philos Mag* 2010;**2**:559–72. doi: 10.1080/14786440109462720.
82. Belarbi MA, *et al*. CA as dimensionality reduction for large-scale image retrieval systems. *Int J Ambient Comput Intell* 2017;**8**:45–58. doi: 10.4018/IJACI.2017100104.
83. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 1998;**20**:832–44. doi: 10.1109/34.709601.
84. Cortes C, Vapnik V. Supprot-vector networks. *Mach Learn* 1995;**297**:273–97. doi: 10.1111/j.1747-0285.2009.00840.x.
85. Huang Y, You Z, Chen X. A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. *Curr Protein Pept Sci* 2018;**19**:468–78. doi: 10.2174/1389203718666161122103057.
86. Meng F, You Z, Chen X, *et al*. Prediction of drug–target interaction networks from the integration of protein sequences and drug chemical structures. *Molecules* 2017;**22**. doi: 10.3390/molecules22071119.
87. Ogata H, Goto S, Sato K, *et al*. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 1999;**27**:29–34. doi: 10.1093/nar/27.1.29.
88. V.M. Khadse *et al*., Statistical study of machine learning algorithms using parametric and non-parametric tests: a comparative analysis and recommendations, *Int J Ambient Comput Intell*. 11 (2020) 80–105. doi: 10.4018/IJACI.2020070105.