AUT Journal of Mathematics and Computing

AUT J. Math. Comput., 6(4) (2025) 297-310
https://doi.org/10.22060/AJMC.2024.23048.1223

Original Article

DRP-VEM: Drug repositioning using voting ensemble model

Zahra Ghorbanali[a], Fatemeh Zare-Mirakabad*[a], Bahram Mohammadpour[a]

[a]*Computational Biology Research Center (CBRC), Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran*

ABSTRACT: Conventional approaches to drug discovery are both expensive and time-intensive. To circumvent these challenges, drug repurposing or repositioning (DR) has emerged as a prevalent strategy. A noteworthy advancement in this field involves the widespread application of machine learning techniques. The effectiveness of these methods depends on the quality of features, their representations, and the underlying dataset. Notably, the issue of redundancy in feature sets can detrimentally impact the overall performance of these methods. Furthermore, the careful selection of a suitable training set plays a pivotal role in enhancing the accuracy of machine learning approaches in addressing drug repurposing challenges. Discovering the appropriate training set faces two significant challenges. Firstly, many methods utilize known drug-disease pairs for positives and unknown pairs for negatives. The stark imbalance in the number of known and unknown pairs often results in a bias towards the larger group, introducing errors in machine learning performance. Secondly, the absence of a documented drug-disease association indicates that it hasn't been experimentally approved yet, and this status may change in the future. This paper introduces DRP-VEM, a novel approach designed for predicting drug repositioning, specifically customized to tackle the challenges previously outlined. DRP-VEM evaluates the effectiveness of binary-based and similarity-based representations of drugs and diseases in enhancing the model's performance. Additionally, it proposes a voting ensemble training strategy, adept at managing imbalanced datasets. The assessment of DRP-VEM spans a range of parameters, including its efficacy in representing both diseases and drugs, the proficiency of its classification methods, and the application of voting ensemble training approaches using heterogeneous evaluation criteria. Significantly, DRP-VEM achieves an AUC-ROC of 81.8% and AUC-PR of 76.6%. Comparative analysis with other studies highlights the superior performance of the proposed model, underscoring its effectiveness in drug repositioning prediction.

Review History:

Received:09 March 2024
Revised:25 May 2024
Accepted:27 May 2024
Available Online:01 October 2025

Keywords:

Drug repurposing
Voting model
Ensemble learning

MSC (2020):

92B20; 68T30; 68U99

1. Introduction

Despite the notable advancements in technology and its significant contributions to disease diagnosis, the translation of these successes into practical medical treatments lags behind [1]. Traditional drug discovery methods, which

*Corresponding author.
E-mail addresses: z_ghorbanali@aut.ac.ir, f.zare@aut.ac.ir, b.moammadpour@aut.ac.ir

are both time-consuming and expensive, further impede the pace of progress. Recent studies reveal that the entire process of identifying a new druggable component, navigating through the testing phases, and finally bringing it to market can span over a decade, incurring costs ranging from $314 million to $2.8 billion [19]. Drug repositioning or drug repurposing (DR) strategy uses an approved drug for a new indication outside its first treatment purpose. A notable historical instance of drug repositioning involves sildenafil, initially developed for the treatment of hypertension but now widely employed for the management of erectile dysfunction [1]. The drug repurposing approach is also applicable for identifying effective treatments for emerging diseases. For instance, researchers repurposed existing antiviral medications, including baloxavir, azvudine, and darunavir, to address the challenges posed by the coronavirus disease during the Covid-19 pandemic [8].

The DR approach is initially discovered opportunistically by physicians. Although leveraging retrospective clinical experiences proves beneficial, they must evaluate a broad spectrum of drugs to identify the proper repurposed candidates. Given the lack of systematic approaches in traditional methods, the proliferation of computational techniques, the wide application of these methods in diverse studies, and the increasing abundance of data on drugs and diseases, researchers favor the application of computational methods to address the DR problem. The computational methods can be divided into three main groups based on the applied data: drug-based, disease-based, and hybrid approaches.

With the abundance of drug-related information, an increasing number of researchers are directing their attention towards drug-based techniques. Notably, Ozsoy et al. [15] have made significant contributions in this domain. Their approach integrates three key drug features: chemical structure, protein interaction, and side effects, aiming to tackle the DR problem. They employ the Pareto dominance technique to identify drug neighbors. Additionally, by utilizing a collaborative filtering recommendation system, they assess the probability of associations between drug-disease pairs [15]. Zeng et al. [21] introduced the DeepDR method, a novel approach that employs random walks to represent nine drug feature networks. These representations are then consolidated through an autoencoder. Subsequently, a variational autoencoder is applied to estimate the probabilities of associations between drug-disease pairs [21]. Chen et al. collected three features of the drug: chemical structure, targets, and side effects. Following the calculation of drug similarities, they suggested a fusion method to merge these similarities, predicting the association probabilities of drug-disease pairs [3].

In tackling the drug repositioning challenges, the significance of disease-related data cannot be overstated. However, due to limited information, researchers have not extensively delved into disease-based methodologies. As a result, these approaches have predominantly focused on specific diseases or therapeutic domains [7]. Chiang and Butte, for instance, devised a methodology where they calculated disease similarities based on shared therapies. Employing a 'guilt by association' approach, they incorporated these disease similarities, leading to the identification of novel drug-disease association pairs [5].

In hybrid methodologies, researchers merge both drug and disease data to assess the likelihood of drug-disease association pairs. Moridi et al. [14] devised a pipeline that adeptly captures drug and disease features, leveraging the power of deep learning methods. Through a non-linear approach, they effectively identified potential candidates for drug-disease associations [14]. Similarly, Xuan et al. introduced DisDrugPred [20], a method that integrates drug similarities, disease similarities, and known drug-disease associations using the non-negative matrix factorization technique. This approach calculates the association probability of drug and disease pairs [20]. Lue et al. [13]took a different approach with their RWHND method, reconstructing a heterogeneous network by integrating data on drugs, drug targets, diseases, and disease genes. Subsequently, they applied a random walk model to identify potential pharmaceutical treatments for a given disease [13].

While researchers have made commendable strides in addressing the DR problem, several challenges persist that warrant further exploration. In the following, we delineate these challenges and propose our approaches to overcome them:

- Retrospective studies in the realm of DR have traditionally leaned towards a focus on drug-based methods, largely sidelining hybrid approaches. Furthermore, these investigations frequently adopt a multitude of drug features without a comprehensive consideration of their relative importance. In our study, we endeavor to examine the necessity and potential redundancy associated with utilizing all features in resolving DR problems. Our objective is to identify whether certain features carry more significance and can substantially enhance the accuracy of machine learning methods dedicated to addressing DR challenges.

- Within the existing literature, a plethora of machine learning techniques has been explored to tackle the DR problem. This article contributes to the understanding that, given a well-defined representation and combination of features, the choice of classification methods has only a marginal impact on the predicted outcomes for DR. The emphasis here is on the meticulous selection and integration of features, highlighting that the effectiveness of predictive models in drug repositioning is more heavily influenced by the thoughtful construction of input variables than by the specific classification algorithm employed.

- Many existing approaches commonly designate known drug-disease pairs as positives and all unknown pairs as negatives. However, creating an appropriate training set encounters two notable challenges. Firstly, the substantial imbalance where the number of known pairs is significantly less than unknown pairs often results in machine learning bias towards the majority group, consequently compromising the method's performance [10]. The second challenge stems from the absence of clinically assessed negative data for drug and disease associations. In response to these challenges, this study proposes a novel algorithm for generating a training set – the voting ensemble training approach.

This paper introduces DRP-VEM, a novel approach designed to address the challenges mentioned through the proposed ideas. The primary innovation of this study lies in the application of the voting ensemble method to address the DR problem. To illustrate the suitability of DRP-VEM's selected drug features, the chosen feature representation for drugs and diseases, the selected machine learning method, and the voting ensemble training approach, we conduct a comparative analysis with the DisDrugPred method [20].

2. Material and methods

This article seeks to achieve several key objectives:

- Identify appropriate feature presentation: Determine which feature representations are most suitable for depicting drugs and diseases in the context of the DR problem,
- Feature impact analysis: Analyze the impact of individual features on solving the DR problem, aiming to discern which features play a more crucial role,
- Evaluate the necessity of drug features: Assess whether all drug features are essential or if their inclusion leads to redundancy in predicting drug-disease associations,
- Optimal classification method: Select the classification method that demonstrates the highest accuracy in addressing the DR problem,
- Introducing voting ensemble training approach: Propose a novel voting ensemble training approach to overcome challenges associated with using unknown drug-disease pairs as the negative set and dealing with imbalanced data in the context of the DR problem.

In the following section, we computationally define the DR problem and provide explanations for the used datasets. Additionally, we clarify the data representations and training approach. Subsequently, we delve into the details of our proposed method.

2.1. Computational drug repositioning problem

In the DR problem, a set of diseases, denoted as $\rho = \{P_1, P_2, \ldots, P_n\}$, and a set of drugs, denoted as $\phi = \{R_1, R_2, \ldots, R_m\}$ are defined. Here, n represents the number of diseases, and m represents the number of drugs. In the mathematical formulation of the DR problem, the primary objective is to determine the existence of a therapeutic association between a disease $P \in \rho$ and a drug $R \in \phi$. If the model predicts that the pair $< P, R >$ has a therapeutic association, the output is represented as one; otherwise, it is zero.

In the DR problem, the following data is provided as input:

- The features of the disease P
- The features of the drug R.

The task is to develop a model that can predict the therapeutic association between diseases and drugs, generating binary outputs (0 or 1) based on the presence or absence of such associations. This mathematical representation forms the foundation for the computational formulation of the DR problem.

2.2. Data sources

To establish a robust foundation for our study, it is essential to compile a dataset of known drug-disease associations and identify pertinent features for both drugs and diseases. Here, we introduce the databases utilized for data extraction.

- Drug-Disease associations: The "repoDB" database [2], specifically designed for drug repositioning, is chosen to gather known drug-disease association pairs. This database serves as a reliable source for validated drug-disease relationships.

- Drug features: Drug-related information, including names, identification, targets, and side effects, is retrieved from "DrugBank" [18]. The target domains of drugs are extracted from "UniProt" [6]. Also, information on side effects is obtained from "SIDER4.1" [11]. Finally, The chemical structures of drugs are collected using "PubChem" [9].
- Disease Feature: The disease similarity values are extracted from the "DincRNA" database [4], utilizing Wang's method [17]. Wang's method is a specific approach employed to measure the similarity between diseases within the context of the DincRNA database. This method likely involves considering various factors, such as shared biological pathways or molecular characteristics, to quantify the similarity between different diseases.

The disease list used in this study is constrained by the DincRNA database, comprising 158 distinct diseases. Additionally, from the available data sources, a subset of 413 drugs has been carefully selected, ensuring that all relevant features for these drugs are available. Table 1 summarizes the number of entries extracted from the various databases.

Table 1: The size of applied databases.

Data	Size
Drugs	413
Diseases	158
Known drug-disease associations	748
Targets	1506
Domains	1070
Side effects	5734
Chemical substructures	881

2.3. Data representation

As mentioned earlier, a primary challenge in the DR problem is determining a suitable representation for drugs and diseases. Commonly, the other studies employ one-hot encoding or similarity measures to portray them. The objective of this paper is to evaluate the effectiveness of each one in addressing the DR problem.

2.3.1. Drug feature representation

We define $\mathcal{F}_R^F = \{\xi_1, \xi_2, \ldots, \xi_{l_F}\}$ as a set of feature components for $F \in \{T, D, S, C\}$, where ξ_i shows i_{th} component of feature F, and l_F denotes the number of feature components. Each feature, shown by T, D, S, and C corresponds to the target, domain, side effect, and chemical structure, respectively. In the following, we introduce two types of drug representation named binary (B) and cosine similarity (C), as follows:

- The binary vector B_R^F with length l_F for drug R is defined based on feature F, as:

$$\forall 1 \leq k \leq l_F : B_R^F[k] = \begin{cases} 1 & \text{Component } \xi_k \text{ is related to Drug } R, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

  The binary vectors B_R^T, B_R^D, B_R^S, and B_R^C to represent drug $R \in \phi$ are computed for target, domain, side effect, and chemical structure features based on Eq.1 respectively. Moreover, a global binary vector, B_R^G, is defined based on concatenation of all drug features as below:

$$B_R^G = B_R^T \cdot B_R^D \cdot B_R^S \cdot B_R^C. \tag{2}$$

  Subsequently, the binary vector B_R^{G-F} is considered when each feature $F \in \{T, D, S, C\}$ is omitted once from the global vector.

- The cosine similarity vector C_R^F with length m is presented for drug R based on feature F as follows:

$$\forall 1 \leq k \leq m, R_k \in \phi : C_R^F[k] = \text{Cosine similarity between } B_R^F \text{ and } B_{R_k}^F. \tag{3}$$

The cosine similarity vectors C_R^T, C_R^D, C_R^S, and C_R^C are computed for target, domain, side effect, and chemical structure features to represent drug $R \in \phi$ based on Eq. (3), respectively. We also define two cosine similarity vectors, C_R^G and C_R^N, based on the concatenation and normalizing of all drug features, as follows, respectively:

$$C_R^G = C_R^T \cdot C_R^D \cdot C_R^S \cdot C_R^C, \qquad C_R^N = Norm(C_R^T + C_R^D + C_R^S + C_R^C). \tag{4}$$

Finally, the cosine similarity vector C_R^{G-F} is considered when each feature $F \in \{T, D, S, C\}$ is omitted once from the C_R^G vector. Thus, we make nineteen different drug feature representations as below:

$$\begin{aligned} U = \{ & B_R^T, B_R^D, B_R^S, B_R^C, B_R^G, B_R^{G-T}, B_R^{G-D}, B_R^{G-S}, B_R^{G-C}, \\ & C_R^T, C_R^D, C_R^S, C_R^C, C_R^G, C_R^{G-T}, C_R^{G-D}, C_R^{G-C}, C_R^{G-S}, C_R^N \}. \end{aligned} \tag{5}$$

2.3.2. Disease feature representation

Here, we define two types of disease representation called Wang vector (W) and one-hot vector (O) as below, $V = \{W, O\}$:

- Wang vector W_P with size n is defined for disease $P \in \rho$ based on Wang's similarity function [17], where:

$$\forall 1 \leq k \leq n, P_k \in \rho, W_P[k] = \text{Wang' s similarity between disease } P \text{ and disease } P_k. \tag{6}$$

- The one-hot vector O_P with length n is defined for disease P as follows:

$$\forall 1 \leq k \leq n, P_k \in \rho : O_P[k] = \begin{cases} 1 & P = P_k \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

2.4. The voting ensemble training approach

A conventional methodology for constructing training and test sets in the DR problem entails categorizing known drug-disease association pairs as the positive set (A) and unknown pairs as the negative set (B). This categorization can be defined as follows:

$$A = \{(P, R) | P \in \rho, R \in \phi, \text{there is a known association between disease } P \text{ and drug } R\}, \tag{8}$$

and

$$B = \{(P, R) | P \in \rho, R \in \phi, \text{there is a no known association between disease } P \text{ and drug } R\}. \tag{9}$$

However, this approach is deemed inappropriate due to the following challenges:

- Imbalance in pair numbers: The number of known pairs is significantly less than the unknown ones. This imbalance can introduce bias into binary classifiers, affecting the method's overall performance [10].
- Clinical assessment absence: The absence of an established association between a drug and disease does not imply a lack of potential association. Rather, it indicates that the clinical assessment of this pair has not been conducted yet.

To tackle the initial challenge, we adopt an under-sampling approach, randomly selecting unknown association pairs at a scale k times that of known association pairs. This selection method, utilized in creating the training set, is referred to as a $one-to-k$ distribution. To overcome the second challenge, we cluster unknown drug-disease pairs based on the $one-to-k$ distribution approach to construct negative training sets. In these sets, the intersection set is empty, and the union set comprises the entire dataset. Let the number of these clusters be n_k. Consequently, the model is trained n_k times on these datasets. During testing, for each sample, we aggregate the responses from the trained models to predict associations. The input parameters for this process include sets A and B, as well as an integer k. The methodology for creating training and test sets for the DR problem is outlined as follows:

1. Positive test set (P_{test}): comprises 10% randomly selected samples from the positive set A.
2. Negative test set (N_{test}): consists of $|P_{\text{test}}|$ randomly selected samples from the negative set B.
3. Positive training set (P_{train}): comprises the remaining samples from set A after excluding those in the positive test set ($P_{\text{train}} = A - P_{\text{test}}$).
4. Negative training sets (N_{train_i}): For each i from 1 to n_k,
   - N_{train_i} is the i_{th} negative training set.

- it includes $k * |P_{\text{train}}|$ randomly chosen samples from set B excluding those in the negative test set (N_{test}) and the union of previous negative training sets ($B - N_{\text{test}} - \bigcup_{j=1}^{i-1} N_{\text{train}_j}$).
- The condition $\bigcap_{i=1}^{n_k} N_{\text{train}_i} = \emptyset$ ensures no overlap among different negative training sets, and $\bigcup_{i=1}^{n_k} N_{\text{train}_i} = B - N_{\text{test}}$ covers the entire set B excluding the negative test set.

5. Test set (ϵ): Defined as the union of the positive test set and the negative test set ($\epsilon = P_{\text{test}} \cup N_{\text{test}}$).
6. Training sets (τ_i): For each i from 1 to n_k, it comprises the positive training set (P_{train}) and the i_{th} negative training set (N_{train_i}).
7. Voting ensemble training set: defined as $T_k = \{\tau_i | 1 \leq i \leq n_k\}$.

For each $k = \{1, 2, 3, 5\}$, we define T_k as the voting ensemble training set, comprising n_k sets for training the classifier. Each $\tau_i \in T_k$ serves as input to the classifier as the training set. Therefore, the classifier is trained n_k times. For each sample from the test set, denoted as $x \in \epsilon$, each trained model predicts the association between the disease and drug pairs. Subsequently, we aggregate the predictions from the n_k models using a voting mechanism. Figure 1 illustrates the proposed voting ensemble training approach when $k = 2$, with the negative training set being twice the size of the positive training set.

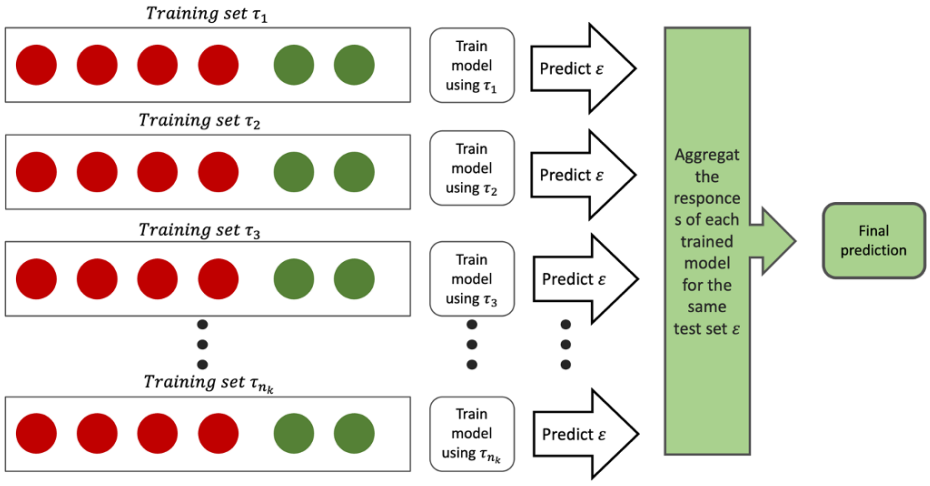

Figure 1: The voting ensemble training approach, when $k = 2$.

2.5. DRP-VEM method

The main steps of the DRP-VEM, comprising of quadruplets such as $model =< v, u, c, t >$, are available in Figure 2 and the details are as follows:

1. Selecting a proper disease representation, $v \in V = \{O, W\}$.
2. Picking an appropriate drug representation,

$$u \in U = \{B_R^T, B_R^D, B_R^S, B_R^C, B_R^G, B_R^{G-T}, B_R^{G-D}, B_R^{G-S}, B_R^{G-C}, C_R^T, C_R^D, C_R^S, C_R^C, C_R^G, C_R^{G-T}, C_R^{G-D}, C_R^{G-C}, C_R^{G-S}, C_R^N\}.$$

3. Concatenating v and u, making $input = v.u$, for feeding into a classifier.
4. Choosing proper classification method among decision tree (DT), random forest (RF), and complement naïve bayes (CNB) classifiers named $c \in C = \{DT, RF, CNB\}$.
5. Electing a voting ensemble training approach called $t \in T = \{T_1, T_2, T_3, T_5\}$.

We assess the different combinations of parameters to find the best model based on our framework.

3. Results and discussion

This section evaluates the possible quadruplets of the form model = $\langle v, u, c, t \rangle$ (refer to Figure 2) and determines the proper disease feature representation $v \in V = \{O, W\}$, drug feature representation

$$u \in U = \{B_R^T, B_R^D, B_R^S, B_R^C, B_R^G, B_R^{G-T}, B_R^{G-D}, B_R^{G-S}, B_R^{G-C}, C_R^T, C_R^D, C_R^S, C_R^C, C_R^G, C_R^{G-T}, C_R^{G-D}, C_R^{G-C}, C_R^{G-S}, C_R^N\},$$

Figure 2: The main steps of DRP-VEM.

classifier method $c \in C = \{DT, RF, CNB\}$, and voting ensemble training approach $t \in T = \{T_1, T_2, T_3, T_5\}$. Therefore, we train 912 different models.

The proposed DRP-VEM model is developed using Python and executed on Google Colab, a cloud-based platform. The hardware specifications of the Colab environment include 12 GB of RAM and 107 GB of disk space. The operating system on which the model is run is a Linux-based OS, which is the default environment provided by Google Colab. In the following, the evaluation criteria are introduced and the results are described. Subsequently, the DRP-VEM and DisDrugPred are compared based on the same training and test sets [20].

3.1. Evaluation Criteria

In our evaluation process, we employ four distinct evaluation criteria to assess each $model_z$, where $1 \leq z \leq 912$, including accuracy (ACC), area under receiver operating characteristic curve ($AUC - ROC$), area under the precision-recall curve ($AUC - PR$), and weighted average score (WAS).

ACC shows the rate of correct prediction to all predictions as below:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}. \tag{10}$$

where the definition of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) is available in Table 2.

Table 2: The definitions of TP, FP, TN, and FN.

Term	Description
TP	The number of known associations that DRP-VEM recognized them correctly.
FP	The number of unknown associations that DRP-VEM recognized as associated wrongly.
TN	The number of unknown associations that DRP-VEM recognized them correctly.
FN	The number of known associations that DRP-VEM recognized as not associated wrongly

$AUC - ROC$ [12] is the area under the receiver operating characteristic curve, which uses different ranking cutoffs and curves the true positive rate (TPR) and false positive rate (FPR) where,

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}. \tag{11}$$

$AUC - PR$ [16] shows the area under a precision-recall curve, a plot of precision and recall, where,

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN}. \tag{12}$$

The WAS score is calculated based on integrating ACC, $AUC - ROC$, and $AUC - PR$ as below:

$$WAS = \frac{2 * (AUC - ROC) + 2 * (AUC - PR) + ACC}{5}. \tag{13}$$

As a result, the set of evaluation criteria is defined as $E = \{ACC, AUC - ROC, AUC - PR, WAS\}$. For each $e \in E$ and $1 \leq z \leq 912$, the evaluation of $\text{model}_z =< v_z, u_z, c_z, t_z >$ on the test set is shown by $e(\text{model}_z)$.

3.2. DRP-VEM analysis based on different parameters

One of the primary objectives of DRP-VEM is to evaluate the effectiveness of selected features and recommend the most crucial ones for further studies, thereby suggesting the appropriate representation for them. Additionally, identifying a useful classifier and determining the optimal number for k in the process of voting ensemble method are another important aspect that we should consider. To accomplish these goals, we assess each quadruplet model separately.

- **Assessment of disease feature representation**

  Two introduced representations for disease are Wang (W, see Eq. (5)) and one-hot (O, see Eq. (6)) vectors. The W vector is derived using Wang similarity measurements, while the O vector is a binary representation determining the given disease. The performance of each $v \in V = \{W, O\}$ for every evaluation criterion $e \in E$ is calculated according to Eq.14. The performance of each representation on an average is available in Table 3.

$$avg(v_z = v) = \frac{1}{|U| \cdot |C| \cdot |T|} \cdot \sum_{u_z \in U} \sum_{c_z \in C} \sum_{t_z \in T} \text{model}(v, u_z, c_z, t_z). \tag{14}$$

Table 3: The average performance of DRP-VEM for selecting the proper disease representation.

Disease representation	ACC(%)	AUC-ROC(%)	AUC-PR(%)	WAS(%)
W	77.2	69.8	51.8	69.16
O	75.8	68.2	50.1	67.62

According to Table 3, the W vector achieves higher performance compared to the O vector. This superiority may be attributed to the W vector incorporating more information about other diseases and their relationships based on Wang similarity. Therefore, it is advisable to use a similarity-based vector when representing diseases.

- **Assessment of drug feature selection and representation**

  When preparing the drugs for input into the model, two main approaches are proposed: binary-based, indicating the presence or absence of a feature, and similarity-based, derived from binary vectors using cosine similarity measurements. Additionally, to establish a baseline case where all features are considered, and to determine the relative importance of each feature, we conduct two types of experiments: one using only one feature at a time and another where each feature is omitted individually. Therefore, 19 different representation are proposed for drugs according to Eq. (5). The performance of each $u \in U$ for every evaluation criterion $e \in E$ is measured using Eq. (15).

$$avg(u_z = u) = \frac{1}{|V| \cdot |C| \cdot |T|} \cdot \sum_{v_z \in V} \sum_{c_z \in C} \sum_{t_z \in T} \text{model}(v_z, u, c_z, t_z). \tag{15}$$

  Figure 3 illustrates the performance of each drug representation proposed in Eq. 5, averaged across all cases according to Eq. 15. The blue, green, orange, and pink charts represent ACC, AUC-ROC, AUC-PR, and WAS criteria, respectively.

  Following the comparison, the cosine similarity representations perform slightly better than the binary representations. This improvement is attributed to the information contained about the relationships between the drugs stored in this representation. Upon analyzing the models, the highest score is obtained when the side effect is employed as the drug feature and cosine similarity as the representation vector. This demonstrates the superior importance of side effects compared to using all features. Furthermore, the results highlight the significance of side effects as the most crucial feature to consider when addressing DR problems. The absence of side effects adversely affects the performance of baseline models, emphasizing that relying solely on this feature yields more accurate results.

- **Assessment of classification method**

  In this study, DT (Decision Tree), RF (Random Forest), and CNB (Complement Naive Bayes) are considered baseline machine learning models, aligning with similar approaches in other research. The advantages of these models can be summarized in various aspects, including handling the non-linearity in the data, suitability for imbalanced data, and Learning in an ensemble way. The performance of each classifier $c \in C$ for every evaluation criterion $e \in E$ is calculated according to Eq. (16). The results are available in Table 4.

$$avg(c_z = c) = \frac{1}{|V| \cdot |U| \cdot |T|} \cdot \sum_{v_z \in V} \sum_{u_z \in U} \sum_{t_z \in T} \text{model}(v_z, u_z, c, t_z). \tag{16}$$

  Based on the results in Table 4, the DT model significantly outperforms the CNB model, followed by the RF model. In terms of instances, it achieves approximately 20% higher ACC and 23% higher AUC-PR than CNB. Moreover, it exhibits approximately 7% higher AUC-ROC and 8% higher AUC-PR than the RF model.

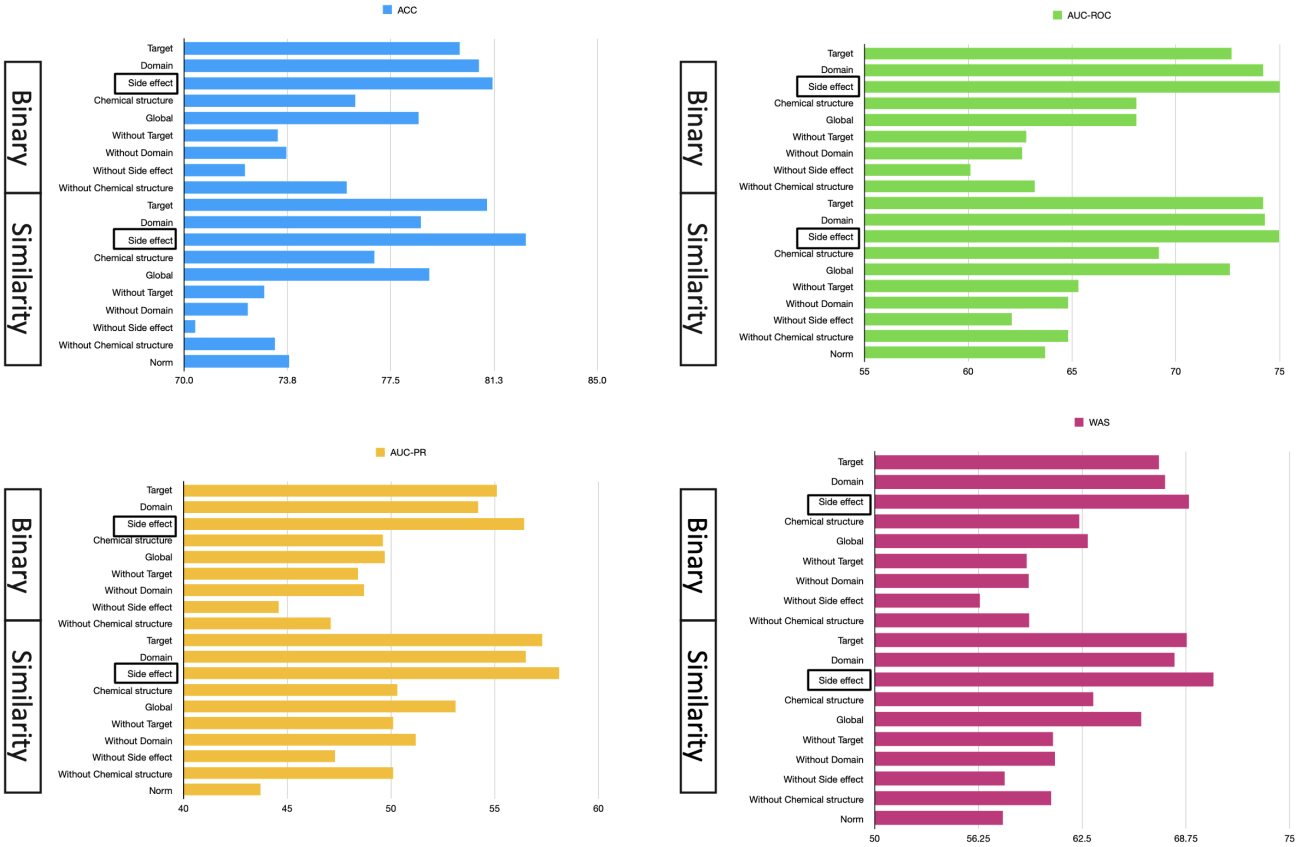

Figure 3: The performance of each drug representation based on different criteria.

Table 4: The performance of each classifier based on the average of the models.

Classifier	ACC(%)	AUC-ROC(%)	AUC-PR(%)	WAS(%)
DT	84.0	77.2	61.6	72.32
RF	80.3	70.0	52.8	65.20
CNB	67.6	61.9	38.7	53.8

- **Assessment of voting ensemble training approach**

  The proposed voting ensemble training method utilizes a one-to-k distribution, dividing the negative training set into subsets with a size k times that of the positive training data. These subsets are then used to train the models, and their predictions for a fixed test set are aggregated in a voting manner to make the final prediction (refer to Figure 1). In this study, k is considered as 1, 2, 3, and 5. Table 5 presents the evaluation criteria $e \in E$ for each $t_z \in T = \{T_1, T_2, T_3, T_5\}$ according to Eq. (17).

$$avg(t_z = t) = \frac{1}{|V| \cdot |U| \cdot |C|} \cdot \sum_{v_z \in V} \sum_{u_z \in U} \sum_{c_z \in C} \text{model}(v_z, u_z, c_z, t). \tag{17}$$

Table 5: The performance of each size of k based on the average of the models.

Classifier	ACC(%)	AUC-ROC(%)	AUC-PR(%)	WAS(%)
T_1	71.1	71.3	66.2	69.2
T_2	75.2	69.6	52.7	64.0
T_3	79.9	70.0	48.3	63.3
T_5	83.1	67.6	36.9	58.4

Based on the findings in Table 5, when the size of positive and negative training sets are equal ($k = 1$),

the models demonstrate superior overall performance. The decline in machine learning model accuracy, particularly with limited data (the positive set), is linked to the increase in k and the expansion of the negative set.

- **Assessment concatenation between different drug and disease features**

  Another aspect requiring assessment is the quality of input representation, which is obtained from the concatenation of drug and disease representations. To analyze this, we concatenate various combinations of drug representations ($u \in U$) and disease representations ($v \in V$) to identify the most suitable ones. We evaluate these combinations based on their performance across all classifiers and voting ensemble training sets, as per Eq. (18). The corresponding results are depicted in Figure 4. The blue, green, orange, and pink charts represent ACC, AUC-ROC, AUC-PR, and WAS criteria, for each drug representation concatenated with a specific disease representation (Wang or one-hot) based on Eq. (18), respectively.

$$avg(v_z = vandu_z = u) = \frac{1}{|C| \cdot |T|} \cdot \sum_{c_z \in C} \sum_{t_z \in T} \text{model}(v, u, c_z, t_z). \tag{18}$$

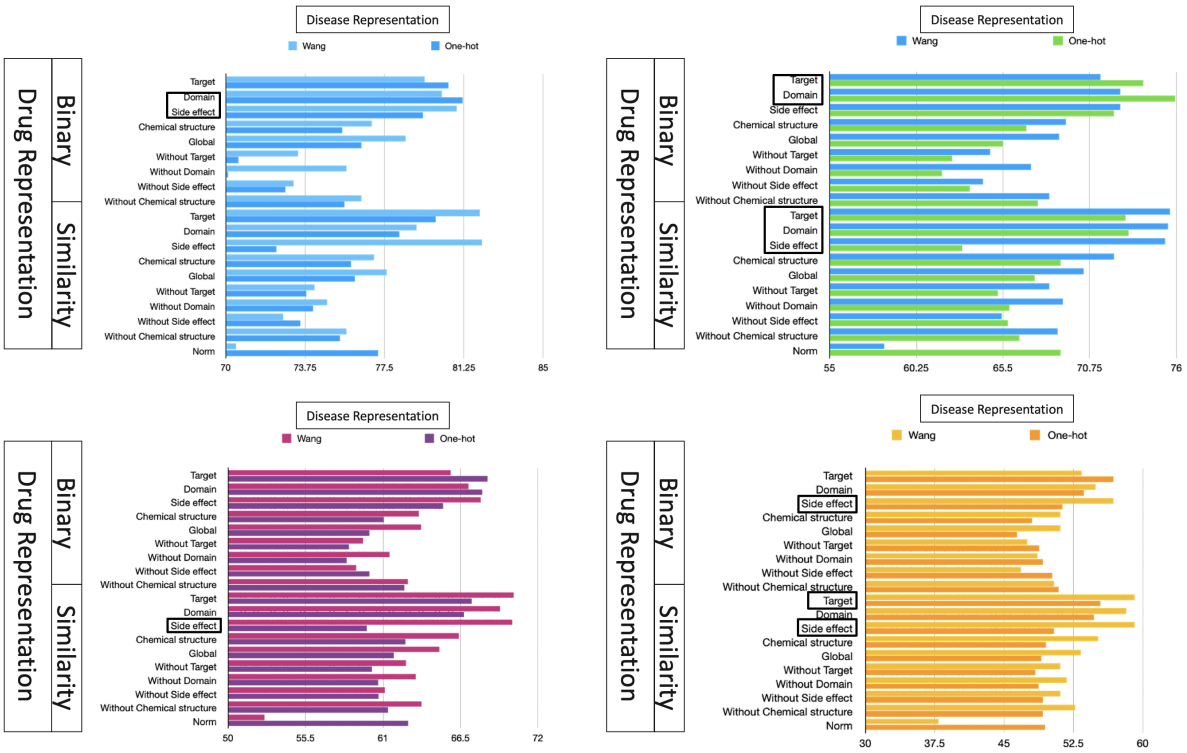

Figure 4: The performance of drug and disease representation concatenation based on different criteria.

According to the findings, the target and side effect features are identified as more crucial than others, as omitting or including them leads to higher model scores. Furthermore, the combination of representations based on similarity and those based on one-hot encoding synergize better than other combinations. Among all trained models, the optimal combination is achieved by concatenating the cosine similarity representation of the target feature for drugs and using the Wang vector for diseases.

- **Assessment of the effectiveness of classification methods**

  This study considers DT, RF, and CNB as the baseline machine learning approaches to propose the more suitable of them in addressing DR problem. However, the quality of classifiers is dependent on the voting ensemble training sets. Therefore, every classifier $c \in C$ uses a voting ensemble training set ($t \in T$) for learning the associations between drug and disease pairs. We analyze all combinations of classifiers and training approaches to declare the best model owing to Eq. (19). The results are shown in Figure 5.

$$avg(c_z = candt_z = t) = \frac{1}{|V| \cdot |U|} \cdot \sum_{v_z \in V} \sum_{u_z \in U} \text{model}(v_z, u_z, c, t). \tag{19}$$

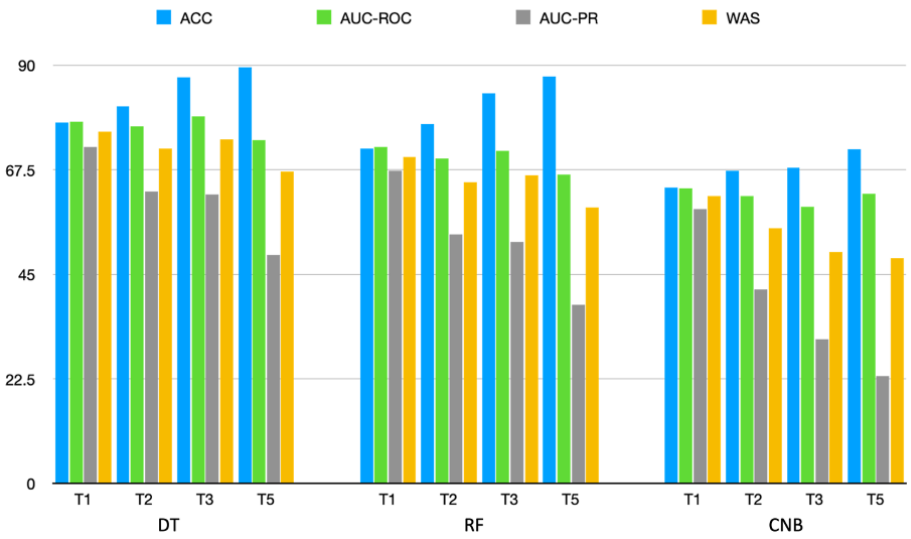

Figure 5: The performance of machine learning methods considering the voting ensemble training approach based on different criteria.

Upon analyzing Figure 5, it is evident that increasing the parameter k in the voting ensemble training approach significantly improves the ACC score of all models. However, this improvement is accompanied by a decrease in the AUC-PR score, reflecting the model's capability to predict accurate positive data. It is crucial to note that in DR problem, positive data is less abundant than negative data, and the goal of the DR problem is to identify potential associations between drugs and diseases. In other words, by constraining the negative set, we can predict positive data more accurately, aligning with our primary goal, as indicated by the AUC-PR score. Based on the results, the most effective classification method involves applying DT as the machine learning method, with $k = 1$ for T_1.

- **Assessment of the input representation impact on model performance**

  While the significance of feature representation and model training approaches has been previously discussed, this section aims to elucidate the impact of feature representation quality on model performance. To this end, we analyze the results of the model for each classifier, utilizing similarity-based and binary-based representations for drugs, alongside Wang and one-hot representations for diseases. Our findings demonstrate that, irrespective of the chosen classifier, alterations in feature representation directly influence model performance. The results indicate that the model achieves optimal performance when the similarity-based representation of drugs is concatenated with the Wang representation of diseases. This combination consistently produces superior results across all models, regardless of the classification method employed. Figure 6 presents stacked bar charts depicting the ACC, $AUC-ROC$, and $AUC-PR$ metrics for each classifier, based on the different drug and disease representations. The green bars represent the binary-based representation of drugs, while the blue bars correspond to the similarity-based representation. For each classifier, there are two columns illustrating the One-hot and Wang representations for diseases.

3.3. Comparing DRP-VEM with DisDrugPred

To assess the effectiveness of DRP-VEM, a comparison is made with another model, DisDrugPred [20], using the same training and test sets. To do so, DisDrugPred [20] is implemented by us, and its evaluation criteria on the test set are obtained. The DisDrugPred algorithm is selected as a comparative benchmark in our study because it represents a state-of-the-art model in the field of drug-disease association prediction. The choice is justified by similar selected features and its reliance on the same set of known drug-disease associations as our proposed method. This alignment ensures a fair and relevant comparison, allowing us to effectively evaluate the performance and advantages of our model against a well-established and widely recognized benchmark.Additionally, both DRP-VEM and DisDrugPred utilize all unknown drug-disease pairs instead of under-sampling, making them directly comparable for assessing whether the matrix-factorization-based model or the newly proposed voting ensemble model demonstrates superior performance. Since DisDrugPred is a regression model and not a classifier, we calculate its mean square error (MSE) instead of ACC.

For DRP-VEM, there are two different cases considered for comparison: BestOverAll and BestAmongAll. The BestOverAll model is derived from the concluded best combination of parameters in the previous subsection. This

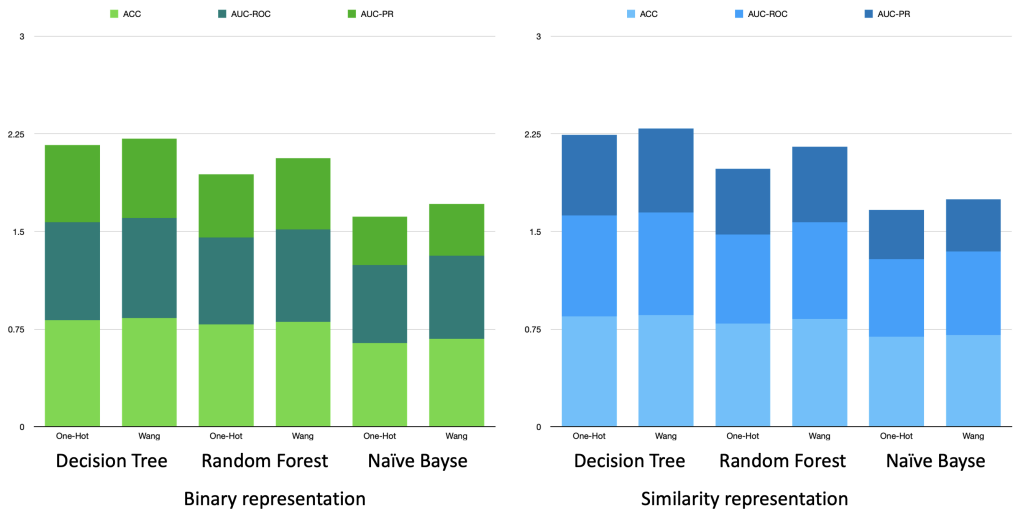

Figure 6: The performance of machine learning methods considering the voting ensemble training approach based on different criteria.

model utilizes the concatenation of the target cosine similarity representation for drugs and the Wang vector for diseases as input, with the aggregation of DT and T_1 as the classification method. However, the best results among all performed models are obtained by using the combination of the side effect binary representation for drugs and the Wang vector for diseases as input. The aggregation of DT and T_3 is used as the classification method for this BestAmongAll model. These two models (BestOverAll and BestAmongAll) are utilized For the comparison between DRP-VEM and DisDrugPred.

As the ACC score is not available for DisDrugPred, and similarly, MSE is not available for DRP-VEM, we calculate the WAS as the average of $AUC-ROC$ and $AUC-PR$. The corresponding results are presented in Table 6. According to the table, the proposed model significantly outperforms DisDrugPred. DisDrugPred, being matrix factorization-based, considers all positive and negative samples at once, and the imbalanced dataset can affect its performance. In contrast, DRP-VEM, through the application of a voting ensemble training approach, effectively addresses the imbalance in the dataset and surpasses DisDrugPred.

Table 6: The performance of each size of k based on the average of the models.

Model	MSE	ACC(%)	AUC-ROC(%)	AUC-PR(%)	WAS(%)
DisDrugPred	0.011	-	82.4	63.9	73.1
BestAmongAll	-	90.2	91.1	85.6	88.3
BestOverAll	-	81.9	81.8	76.6	79.7

4. Conclusion

This article introduces a novel framework, DPR-VEM, designed to address the DR problem through the utilization of a voting ensemble method. The research explores various parameters to identify the optimal combination of drug and disease feature representations, classification methods, and training approaches. The evaluation of the framework is based on key metrics, including Accuracy (ACC), Area Under Receiver Operating Characteristic Curve ($AUC-ROC$), Area Under the Precision-Recall Curve ($AUC-PR$), and Weighted Average Score (WAS).

After an extensive analysis, the results reveal that the Best Overall model within the proposed framework is characterized by the use of target cosine similarity vectors for drug feature representation, Wang similarity vectors for disease representation, a decision tree as the classification model, and a one-to-one distribution for the voting ensemble training approach. The performance metrics for the BestOverall model are as follows: $ACC = 81.9\%$, $AUC-ROC = 81.8\%$, $AUC-PR = 76.6\%$, and $WAS = 79.7\%$. Furthermore, a comparative analysis is conducted between DPR-VEM and DisDrugPred, a state-of-the-art drug repositioning method. The results indicate that DPR-VEM outperforms DisDrugPred, with $AUC-ROC = 82.4\%$, $AUC-PR = 63.9\%$, and $WAS = 73.1\%$ for DisDrugPred.

In conclusion, this study highlights several key findings in the context of DR. Notably:

- Significance of target and domain as drug features: Target or domain information emerges as crucial for effective drug feature representation, emphasizing the importance of considering specific aspects of drugs in the model
- Reduced accuracy and redundancy with concatenation of all features: concatenating all features diminishes model accuracy and introduces redundancy, underlining the necessity of discerning the most informative features for drug representation
- Accuracy of cosine similarity Wang vector as feature representations: The use of cosine similarity vectors and Wang vectors for drug and disease feature representation, respectively, demonstrates higher accuracy in model performance
- Effectiveness of decision tree classifier: The decision tree classifier proves to be more effective in distinguishing the dataset compared to other classification methods Mitigation of biasing challenges with voting ensemble approach: The application of the voting ensemble approach addresses challenges related to classification method biasing, contributing to a more balanced and robust model.

While the focus of this article was primarily on the assessment of drug feature representation, future research endeavors aim to delve deeper into the analysis of disease feature representations. Additionally, the study utilized fingerprint as a representation of drug chemical structure, and while SMILES representation appeared more informative, efforts are directed towards achieving a standardized representation format for SMILES in subsequent research. These in-sights contribute to the refinement and advancement of methodologies in the field of drug repositioning.

References

[1] T. T. Ashburn and K. B. Thor, *Drug repositioning: identifying and developing new uses for existing drugs*, Nature Reviews Drug Discovery, 3 (2004), pp. 673–683.

[2] A. S. Brown and C. J. Patel, *A standard database for drug repositioning*, Scientific Data, 4 (2017), p. 170029.

[3] H. Chen, Z. Zhang, and J. Zhang, *In silico drug repositioning based on the integration of chemical, genomic and pharmacological spaces*, BMC Bioinformatics, 22 (2021), p. 52.

[4] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, *Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function*, Bioinformatics, 34 (2018), pp. 1953–1956.

[5] A. Chiang and A. Butte, *Systematic evaluation of drug–disease relationships to identify leads for novel drug uses*, Clinical Pharmacology & Therapeutics, 86 (2009), pp. 507–510.

[6] T. U. Consortium, *Uniprot: the universal protein knowledgebase in 2021*, Nucleic Acids Research, 49 (2020), pp. D480–D489.

[7] J. T. Dudley, T. Deshpande, and A. J. Butte, *Exploiting drug–disease relationships for computational drug repositioning*, Briefings in Bioinformatics, 12 (2011), pp. 303–311.

[8] C. Harrison, *Coronavirus puts drug repurposing on the fast track*, Nature biotechnology, 38 (2020), pp. 379–381.

[9] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, *Pubchem 2019 update: improved access to chemical data*, Nucleic Acids Research, 47 (2018), pp. D1102–D1109.

[10] B. Krawczyk, *Learning from imbalanced data: open challenges and future directions*, Progress in Artificial Intelligence, 5 (2016), pp. 221–232.

[11] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, *The sider database of drugs and side effects*, Nucleic Acids Research, 44 (2015), pp. D1075–D1079.

[12] R. Kumar and A. Indrayan, *Receiver operating characteristic (roc) curve for medical researchers*, Indian Pediatrics, 48 (2011), pp. 277–287.

[13] H. Luo, J. Wang, M. Li, J. Luo, P. Ni, K. Zhao, F.-X. Wu, and Y. Pan, *Computational drug repositioning with random walk on a heterogeneous network*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16 (2019), pp. 1890–1900.

[14] M. Moridi, M. Ghadirinia, A. Sharifi-Zarchi, and F. Zare-Mirakabad, *The assessment of efficient representation of drug features using deep learning for drug repositioning*, BMC Bioinformatics, 20 (2019), p. 577.

[15] M. G. Ozsoy, T. Özyer, F. Polat, and R. Alhajj, *Realizing drug repositioning by adapting a recommendation system to handle the process*, BMC Bioinformatics, 19 (2018), p. 136.

[16] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, *The area under the precision-recall curve as a performance metric for rare binary events*, Methods in Ecology and Evolution, 10 (2019), pp. 565–577.

[17] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, *Inferring the human microrna functional similarity and functional network based on microrna-associated diseases*, Bioinformatics, 26 (2010), pp. 1644–1650.

[18] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, *Drugbank 5.0: a major update to the drugbank database for 2018*, Nucleic Acids Research, 46 (2017), pp. D1074–D1082.

[19] O. J. Wouters, M. McKee, and J. Luyten, *Estimated research and development investment needed to bring a new medicine to market, 2009-2018*, JAMA, 323 (2020), pp. 844–853.

[20] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, and T. Shen, *Drug repositioning through integration of prior knowledge and projections of drugs and diseases*, Bioinformatics, 35 (2019), pp. 4108–4119.

[21] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, *deepdr: a network-based deep learning approach to in silico drug repositioning*, Bioinformatics, 35 (2019), pp. 5191–5198.

Please cite this article using:

Zahra Ghorbanali, Fatemeh Zare-Mirakabad, Bahram Mohammadpour, DRP-VEM: Drug repositioning using voting ensemble model, AUT J. Math. Comput., 6(4) (2025) 297-310
https://doi.org/10.22060/AJMC.2024.23048.1223