Bioinformatics and Biology Insights
Volume 19: 1–14
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322251328269
journals.sagepub.com/home/bbi
Sage

Bayesian Inference for Drug Discovery by High Negative Samples and Oversampling

Manh Hung Le[1], Nam Anh Dao[1] and Xuan Tho Dang[2]

Abstract

Drug repositioning holds great promise for reducing the time and cost associated with traditional drug discovery, but it faces significant challenges related to data imbalance and noise in negative samples. In this article, we introduce a novel method leveraging high negative oversampling (HNO) to address these challenges. Our approach integrates HNO with advanced techniques such as network-based graph mining, matrix factorization, and Bayesian inference, specifically designed for imbalanced data scenarios. Constructing high-quality negative samples is crucial to mitigate the detrimental effects of noisy negative data and enhance model performance. Experimental results demonstrate the efficacy of our approach in enhancing the performance of drug discovery models by effectively managing data imbalance and refining the selection of negative samples. This methodology provides a robust framework for improving drug repositioning, with potential applications in broader biomedical domains.

Keywords

Drug-disease associations, over-sample, protein associations, imbalanced data, drug repositioning, Bayesian inference

Received: 2 October 2024; accepted: 3 March 2025

Introduction

In the context of today's increasingly complex diseases and rapidly mutating variants posing significant risks to humans, there is an urgent need for treatment methods and drugs that can swiftly respond to these health challenges. Meanwhile, the traditional process of drug discovery and development is becoming less appealing due to its considerable costs and time requirements.[1] Bringing a new drug to market typically spans over a decade and may incur costs amounting to several billion US dollars.[2] In this context, drug repurposing, the process of finding new applications for already approved drugs, emerges as a promising strategy to optimize investment in drug development. Supported by readily available data on drug safety and efficacy, drug repurposing can significantly reduce the time and resources compared to developing a new drug from scratch.[3] A crucial initial step in this process is the selection of a candidate drug and the identification of a new indication for it before proceeding to preclinical trials. To enhance the efficiency of this process, various computational methods have been developed, employing different strategies. Among the most common are similarity-based techniques, which infer new indications based on the similarity of drugs and suggest treatment methods for similar disease conditions. To enhance the efficiency of this process, various computational methods have been developed, employing different strategies. Among the most common are similarity-based techniques, which infer new indications by analyzing the similarity of drugs and suggesting treatment methods for related disease conditions. Recent advancements in computational drug discovery highlight the potential of advanced machine learning techniques. For instance, TransEDRP[4] employs a dual-transformer framework to integrate chemical and pharmacological properties, achieving a 22.67% improvement in accuracy across multiple datasets. Similarly, MilGNet,[5] using heterogeneous graph neural networks, enhances drug-disease association prediction and outperforms 10 state-of-the-art methods. These approaches have also been validated through real-world applications, such as identifying Methotrexate for mismatch repair cancer syndrome. Furthermore, Gottlieb et al introduced PREDICT,[6] a method leveraging drug-drug and disease-disease similarities based on their characteristics to uncover new drug structures and disease associations from phenotypic data.

In their latest study, Yu et al[7] employed similarity analysis to explore the relationships between drugs, based on their 5

[1]Electric Power University, Hanoi, Viet Nam
[2]Academy of Policy and Development, Hanoi, Viet Nam

Corresponding Author:
Xuan Tho Dang, Academy of Policy and Development, Hanoi, Viet Nam.
Email: thodx@apd.edu.vn

principal attributes, as well as the similarities between diseases. This effort culminated in the development of a novel technique, layer attention graph convolutional network (LAGCN), aimed at drug repositioning. By integrating diverse data, this method not only enhances the reliability of identifying similarities between drugs and diseases but also broadens the potential to discover new links between them. However, as highlighted in prior studies, the number of drug-disease relationships experimentally confirmed to date remains limited compared to the total potential relationships.[8] In this context, clearly distinguishing between "positive" (known) and "negative" (unidentified or non-existent) drug-disease pairs emerges as a significant challenge, especially in constructing an effective "negative" dataset to improve the accuracy of machine learning models.

For instance, in the study by Li et al,[9] the number of drugs was 2593, and the number of diseases was 19941, but known drug-disease associations represent only a very small proportion of the 50 million possible drug-disease associations. In machine learning tasks, these known relationships are labeled as positive. However, identifying and pinpointing the high negative drug-disease pairs among the remaining 50 million possible pairs presents a significant challenge, as both potential positive and negative elements are hidden within. In this study, we propose a method to efficiently construct a highly negative dataset, clearly distinguishing it from the known positive elements, thereby enhancing the accuracy of the model.[3,10] In addition, a major challenge identified is the imbalanced data, as observed in the example above, where the positive elements account for only 0.34%. This imbalance affects the precision and reliability of predictions in machine learning models. Hence, this study aims to address the issue of data imbalance and offers specific solutions to improve the quality of the drug repositioning process.

In this study, we endeavor to address the issues mentioned above and make specific contributions toward this end as follows: (a) We analyze the challenge of drug repositioning, predicting new drug-disease relationships, the necessity of constructing a robust high negative dataset of unparalleled quality, and the imbalanced data issue encountered in resolving this challenge by Bayesian inference. (b) We propose a method for constructing a robust high negative dataset of unparalleled quality and apply imbalance techniques for the dataset. The advantages and disadvantages of current methodologies for addressing imbalanced data are investigated and assessed. (c) Through experimental results, we demonstrate that our proposed approach, which combines the creation of a high-quality negative dataset with methods to over-sample, effectively resolves the issue of data imbalance, thereby enhancing the efficiency and reliability of identifying new drug-disease relationships.

Related work

Drug repositioning, which seeks new therapeutic uses for existing drugs, offers a strategy that is both time and cost-efficient due to pre-existing knowledge of drugs' pharmacological and safety profiles.[11] Success stories, like sildenafil (Viagra) for erectile dysfunction and minoxidil for hair loss, are uncommon and typically accidental.[12] The process is hindered by its dependence on prior knowledge and the prohibitive cost of clinical trials, making wide-scale application challenging. Fifty million potential drug-disease pairings manually is infeasible, highlighting the need for computational methods to discover new drug applications efficiently.

The rapid increase in drug and disease-related data, along with advancements in machine learning, has spurred the creation of diverse theoretical methods. These are designed to reveal new drug uses by identifying potential drug-disease associations. These computational strategies are primarily categorized into 3 principal approaches: drug-based, disease-based, and network-based. Each category adopts a distinct premise for association prediction, leveraging the intrinsic properties and similarities within drugs and diseases to facilitate the discovery of novel therapeutic applications.

Drug-based and disease-based methodologies operate on the foundational hypothesis that drugs with similar structural or functional traits tend to be effective against diseases sharing analogous pathogenic processes or symptoms. This concept has been supported by numerous studies, establishing a solid theoretical basis for these approaches. In this context, Wang et al[13] developed a support vector machine (SVM) model to detect potential drug-disease interactions. This model integrates a wide range of data, including molecular structure, molecular activity, and phenotype information, thereby substantially improving the accuracy of therapeutic connection predictions. Echoing this approach, Khalid and Sezerman[14] presented an integrative method that uses a similarity-based framework to predict approved and novel drug targets and their new disease associations. This method integrates protein-protein interactions (PPI), biological pathways, binding site structural similarities, and disease-disease similarity metrics to enhance prediction accuracy. Further contributing to this field, Zhang et al[15] unveiled a novel Similarity Constrained Matrix Factorization for Drug-Disease Association (SCMFDD) prediction aimed at elucidating the connections between drugs and diseases. Leveraging existing drug-disease associations along with drug characteristics and disease semantic data, the SCMFDD projects these relationships into bidimensional spaces to uncover latent features of both drugs and diseases.

The last one is based on the principle of "guilt-by-association" that drugs treating the same disease share structure/network properties and the diseases treated with the same drug also share phenotype/network properties. In their seminal work, Yang et al[16] constructed 3 causal networks targeting cardiovascular diseases, diabetes mellitus, and neoplasms using a causal inference-probabilistic matrix factorization (CIPMF) methodology. This approach aimed to predict and classify drug-disease associations, thereby aiding in the identification of new drug repositioning opportunities. It entailed the integration of multilevel systematic relationships between drugs and diseases from diverse databases to establish causal networks that link drug-target-pathway-gene-disease. Liu et al,[17] in a related vein, developed a heterogeneous network comprising drug-drug similarity, disease-disease similarity, and known drug-disease association networks. They introduced a novel 2-pass random walk with a restart algorithm to predict novel indications for approved drugs. Zhang et al[18] further advanced the domain by formulating drug-disease associations as a bipartite network and implementing a network topological similarity-based inference method, which leverages linear neighborhood similarity to predict unobserved drug-disease associations. Building on these foundational studies, Yue et al[19] conducted a rigorous evaluation of 11 distinct graph embedding methodologies across 3 critical biomedical link prediction tasks: drug-disease association (DDA), drug-drug interaction (DDI), and PPI predictions. Their analysis extended to 2 node classification tasks, specifically the classification of medical term semantic types and the prediction of protein functions. This comprehensive assessment seeks to shed light on the efficacy and practicality of graph embedding techniques in biomedical research, setting a new benchmark for future investigations in this rapidly evolving field.

Within the current landscape of methodologies, 2 principal challenges are prevalent: First, supervised learning-based approaches typically require both positive and negative samples for training predictive models. However, these methods often operate under the assumption that unknown drug-disease pairs default to the negative class, leading to a scarcity of experimentally validated negative samples. This assumption inaccurately constructs the negative sample set, as these unknown pairs could potentially belong to an undefined category, being either positive or negative. Therefore, there is a critical need for a methodology capable of identifying robust negative elements distinctly from known drug-disease pairs. To address the first challenge of constructing robust negative samples, prior studies have proposed various approaches to enhance the reliability of the negative sample set. Methods like EMP-SVD[20] and TS-SVD[21] enhance negative dataset reliability by excluding pairs sharing common proteins or short-path associations within heterogeneous networks, reducing noise from arbitrary assumptions. These methods improve the quality of negative datasets and contribute to the reliability of downstream predictions.

The second challenge arises from the inherent data imbalance prevalent in drug-disease pair datasets, as is common in biological data, where the proportion of positive class elements is substantially lower than that of the negative class. This imbalance significantly affects the predictive efficacy of models. Thus, a method that effectively addresses this imbalance is essential. In this article, we propose an approach that not only constructs a robust set of negative class elements but also tackles the issue of data imbalance, thereby enhancing the predictive accuracy of the model.

Method

Drug repositioning in Bayesian inference

In what follows, the notation d, t, and s refers to the drug, protein, and disease, which are our study objects for drug repositioning based on a comprehensive Bayesian inference.[22] The drug repositioning is expected to provide new drug candidate d for a disease s (equation (1)), and the Bayesian inference offers a classic schema to figure out how drug candidates can be proposed from data analysis:

$$p(d \mid s). \tag{1}$$

The drug-disease prediction process is modeled as a graph mining on a heterogeneous network where nodes are drugs, proteins, and diseases. Importantly, protein examination in drug repositioning should take into account how a drug interacts with a protein and how a protein is related to a disease. Diseases are often caused by mutations involving the binding interface or directing to biochemically dysfunctional allosteric changes in proteins.[23] Considering the conventional association of a drug, a protein, and a disease, the probability of these objects $p(d,t,s)$ can be split into conditional probabilities $p(d \mid t)$ and $p(t \mid s)$ with prior probability $p(s)$ by equation (2):

$$p(d,t,s) = p(d \mid t)p(t \mid s)p(s). \tag{2}$$

For instance, the original probability $p(d \mid s)$ from equation (1) is represented by $p(d,s)\,/\,p(s)$ while $p(d,s)$ can be evaluated by the relation of all known proteins t by equation (3):

$$p(d \mid s) = p(d,\, s)\,/\,p(s) = \sum_t p(d,\, t, s)\,/\,p(s). \tag{3}$$

Since the conventional association of a drug, a protein, and a disease $p(d,t,s)$ in equation (3) was mentioned in equation (3), it is not surprising that the conventional probability of a drug and a disease may be seen in association with proteins:

$$p(d \mid s) = \sum_t p(d \mid t)p(t \mid s)\,/\,p(s). \tag{4}$$

Clearly, such Bayesian inference allows us to include proteins in drug repositioning showing how a new drug can be proposed for a disease through data of confirmed interactions between drugs, proteins, and diseases.

High negative samples and oversampling

Indeed, the learning process for drug repositioning requires training data that consists of pairs of drug d and disease s and their labels saying negative or positive interact based on published documents for each pair. For a given training data Z, the note Z^+ is granted for a set of positive samples and the note of Z^- is made for a set of negative samples:

$$Z = Z^+ \cup Z^-. \tag{5}$$

Identification of the positive drug-disease relation is realized by checking approved drugs from pharmacy companies:

$$Z^+ = \{(d,s), p(d\,|\,s) = 1\}. \qquad (6)$$

The negative drug-disease relation related to those drug-disease whose information of appointment is lacking:

$$Z^- = \{(d,s), p(d\,|\,s) = 0. \qquad (7)$$

There is a specific imbalance in the training data. For any drug d, only a few diseases s are ready for positive samples $p(d\,|\,s) = 1$ while other diseases are set in negative samples:

$$count(Z^+) \ll count(Z^-). \qquad (8)$$

A classification where the data set has skewed class proportions is called imbalanced. Classes that have a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes. The imbalance may cause learning issues. If the number of positive samples is too small relative to negative samples, the training process will spend most of its time on negative samples and positive samples are not learned enough. The above issue leads us to apply a solution as follows.

The finding that the prior probability of protein $p(t)$ may recovered from cross-checking protein-drug association, and the search in training yields inference for protein $p^d(t)$ with the mark of d:

$$p^d(t) = \sum_d p(t\,|\,d)p(d\,|\,t). \qquad (9)$$

Similarly, there can be other way in identifying the prior probability of protein $p(t)$ through protein-disease association, we mark it by $p^s(t)$:

$$p^s(t) = \sum_s p(t\,|\,s)p(s\,|\,t). \qquad (10)$$

Hence, the prior probabilities are joined in a general view from the drug prospect as well as from the disease prospect. This produces the final probability for protein $p(t)$:

$$p^*(t) = p^d(t)p^s(t) = \sum_d p(t\,|\,d)p(d\,|\,t)\sum_s p(t\,|\,s)p(s\,|\,t). \qquad (11)$$

For improving the quality of learning, the negative samples for training can be selected from the set of samples where their probability $p(d\,|\,s) = 0$. Note that the calculation of the $p^*(d\,|\,s)$ is proposed to use equation (4) with the protein prior $p^*(t)$ predefined above by equation (11):

$$p^*(d\,|\,s) = \sum_t p(d\,|\,t)p^*(t)p(t\,|\,s)\,/\,p(s). \qquad (12)$$

The training set of negative samples by equation (12) is likely stronger than the original one due to the sample selection already described:

$$Z_*^- = (d,s), p^*(d\,|\,s) = 0\} \qquad (13)$$

We are thus depicting that selectively constructing training data with high potential negative samples data is possible through evaluation of prior probability for protein, completing data preparation task with the set of samples for training:

$$Z = Z^+ \cup Z_*^- \qquad (14)$$

Prediction in drug repositioning

We have presented the results of training data performed as a part of a balancing number of samples for classes in relation to drug and disease $p(d\,|\,s)$, which covered a range of protein studies. The extensible Bayesian inference by equation (4) allows the prediction of new drug-disease interaction relation. Here, the prior probability of disease $p(s^*)$ is estimated from the training data given a particular disease s^*:

$$p(s^*) = \sum_{(d,s)\in Z} p(s^*\,|\,d)p(d). \qquad (15)$$

The Bayesian inference (equation (4)), therefore, shows a prediction for interaction between a drug d and a disease s which is not presented in the training data Z. It is important to note that our prediction involves voting from 5 predictions, the first one uses the training data prepared from the data set $Z_1 = Z$ (equation (14)) with labels for each pair of drug and disease, and applies the Bayesian rule where the right part uses the training data Z_1:

$$p_1(d\,|\,s) = p(s\,|\,d)p(d)\,/\,p(s),\ Z_1 = Z. \qquad (16)$$

There is thus a Bayesian inference from the above training data, although deeper, using reference of protein in probabilities of $p(d\,|\,t)$ and $p(t\,|\,s)$. This is the second training data Z_2 and exports probability p_2 for test data of pairs (d,s):

$$p_2(d\,|\,s),\ Z_2 = \{\left(d^*\,|\,s^*\right)p(d^*\,|\,s^*) = \sum_t p\left(d^*|t\right)p\left(t|s^*\right)p\left(s^*\right)\}. \qquad (17)$$

After passing through a relation of drug and protein $p(d\,|\,t)$, joined with the relation of protein and disease $p(t\,|\,s)$, it shows other views of data and exports the third prediction as a result:

$$p_3(d\,|\,s),\ Z_3 = \{\left(d^*,s^*\right), p(d^*\,|\,s^*) = \sum_t p(d^*\,|\,t)\sum_{d\in Z} p\left(t|d\right)p\left(d\,|\,s^*\right)\}. \qquad (18)$$

With appropriate inference through drug-disease-drug and again drug-disease, the probability $p(d \mid s)$ is used 2 times with the assistance of $p(s \mid d)$, the learning inference also produces a particular prediction, noted by p_4:

$$p_4(d \mid s),\ Z_4 = \{ \left(d^*, s^*\right), p(d^* \mid s^*) = \sum_{s \in Z} p(d^* \mid s) \sum_{s \in Z} \sum_{d \in Z} p(s \mid d) p(d \mid s^*) \}. \tag{19}$$

It is then necessary to account for the protein with the disease in relation $p(s \mid t)$ and $p(t \mid s)$ yielding the fifth prediction:

$$p_5(d \mid s),\ Z_5 = \{ \left(d^*, s^*\right), p(d^* \mid s^*) = \sum_{s \in Z} p(d^* \mid s) \sum_{s \in Z} \sum_{t} p(s \mid t) p(t \mid s^*) \}. \tag{20}$$

The aim of learning is to maximize belief of reasoning and so get the approximate posterior as close as possible to the true posterior.

To obtain a final prediction for test data with pairs of (d and s), we apply voting of the abovementioned 5 predictions:

$$p(d \mid s) = max_{i=1,..5} p_i(d \mid s). \tag{21}$$

Results and discussion

During the experimental phase, we strictly followed the described method. Initially, we introduced and conducted a detailed analysis of the dataset used in our research to better understand the interrelationships among its components, particularly focusing on drug-disease and protein associations. We then demonstrated that drug discovery using high negative sampling combined with oversampling techniques such as Gaussian-synthetic minority oversampling technique (SMOTE)[24] yields consistent and reliable results.

Data analysis

Data integrity plays a critical role and is rigorously evaluated before being used in experimental design setups. In this research, we used 2 datasets that consist of the dataset introduced by Wu et al[20] and the B-dataset introduced by Zhang et al[15] and Zhao et al.[25] Wu et al[20] constructed their dataset from 3 specific components: drug-protein-disease interactions. Specifically, the disease-protein data was extracted from OMIM,[26] the drug-protein data was sourced from DrugBank,[27] and the drug-disease data was retrieved from Gottlieb's research.[28] Table 1 provides an overview of the disease-protein data sources, illustrating the relationships between 449 diseases and 1467 proteins. Notably, there were 1365 verified drug-disease interactions (considered positive samples), compared with 657 318 unverified interactions (negative samples), resulting in a positive-to-negative sample ratio of 0.207%. Similarly, the drug-protein data included 1186 drugs and 1467 proteins, with 4642 positive samples and 17 352 200 unverified samples, translating to a ratio of 0.267%. The drug-disease data encompassed 1186 drugs and 449 diseases, with 1827 positive samples versus 530 687 unverified samples, achieving a rate of 0.344%. In this study, we proceeded to select high-probability negative samples using formulas previously analyzed.[29]

Table 1. Description of the experimental dataset.

Relation	Number of interactions	Number of no interactions	Ratio in %
Disease-protein (449 × 1147)	1365	657 318	0.207
Drug-protein (1186 × 1147)	4642	1 735 220	0.267
Drug-disease (1186 × 449)	1827	530 687	0.344

The B-dataset includes 269 drugs, 598 diseases, and 1021 proteins. It contains 18 416 drug-disease associations, 3110 drug-protein associations, and 5898 disease-protein associations, as detailed in Table 2. Within this dataset, we constructed positive and negative samples in a manner similar to the one previously introduced. In this article, the matrix $A_{ds}[n \times m]$ represents the drug-disease dataset, the matrix $A_{dp}[n \times k]$ is the drug-protein dataset, and the matrix $A_{sp}[m \times k]$ notes the disease-protein dataset.

Algorithm I : High negative samples selection

```
Input : $A_{dp}\left[n\, x\, k\right]\left(drug - protein - matrix\right)$
        $A_{dp}[m \times k]\ (disease - protei - matrix)$
        $A_{pp}\left[k \times k\right]\left(protein - protein - matrix\right)$
Output : $Z^-$ (Nagative samples set)
Begin Algorithm
$A_{ds} \leftarrow A_{dp} A_{pp} A_{sp}$, See equation (11)
$Z^- \leftarrow \emptyset$;
while $i \le m$ do
    while $j \le n$ do
        if $A(i, j) = 0$ then
            $Z^- \leftarrow Z^- \cup \{(i, j)\}$;
        end
        $j \leftarrow j + 1$;
    end
    $i \leftarrow i + 1$;
end
Return $Z^-$;
End Algorithm
```

Here, n is the number of drugs, m is the number of diseases, and k is the number of proteins. As introduced earlier, the positive samples in the experimental dataset were identified using equation (12), while the status of the remaining samples remained undetermined, they could be either positive or negative.

According to equations (9) to (13), we extract high negative samples (HNS) using Algorithm 1. Ultimately, a new

Table 2. Description of experimental B-dataset.

Relation	Number of interactions	Number of no interactions	Ratio in %
Disease-protein (558 × 1021)	5898	569718	1.04
Drug-protein (269 × 1021)	3110	274649	1.13
Drug-disease (269 × 598)	18416	160862	11.45

dataset was generated by combining the positive samples with these high-quality negative samples according to equation (15); this new dataset is called HNdataset. In addition, we constructed another dataset based on a method for filter negative samples (FNS) introduced by Wu et al,[20] whose name is Fndataset.

Drug repositioning with heterogeneous network

A heterogeneous drug-protein-disease network was constructed, where each node represents a drug d_i, a disease s_j, or a protein p_t. There is an edge between nodes d_i and s_j, d_i and p_t, or s_j and p_t if a previously established relationship exists between them. Different relationships between a drug and a disease s_j are identified through various paths, starting from the drug and ending at the disease. In this study, we used 5 meta-paths introduced[20] with 5 matrices $M1, M2, M3, M4$, and $M5$. M1 represents the relationship between a drug and a disease through the first prediction (equation (16)):

$$M1 = A_{ds}. \tag{22}$$

$M2$ shows the relationship between a drug and a disease through the second prediction (equation (17)):

$$M2 = A_{dp} \times A_{sp}^{T}. \tag{23}$$

$M3$ is based on the third prediction (equation (18)):

$$M3 = A_{dp} \times A_{dp}^{T} \times A_{ds}. \tag{24}$$

$M4$ is used the fourth prediction (equation (19)):

$$M4 = A_{ds} \times A_{ds}^{T} \times A_{ds}. \tag{25}$$

$M5$ represents the fifth prediction (equation (20)):

$$M5 = A_{ds} \times A_{sp} \times A_{sp}^{T}. \tag{26}$$

Extraction of drug-disease features with singular value decomposition

The drug-disease matrices $M1, M2, M3, M4$, and $M5$ collectively voted matrix M by equation (21), which includes 1186 drugs represented in rows and 449 diseases represented in columns. The matrix M has very high dimensionality, making it challenging to analyze and construct models. Singular value decomposition (SVD) is used to reduce the dimensionality of this matrix while retaining essential relationships. This method decomposes the matrix M into 3 component matrices: $\uplus, \Sigma\, and\, \vee$, as follows:

$$M \approx \uplus_{n\times r} \times \Sigma_{r\times r} \times \vee_{r\times m}^{T}.$$

where U contains the left singular vectors. Σ is the diagonal matrix containing singular values, and V contains the right singular vectors.

High negative samples and oversampling

To evaluate and compare efficiency, we categorized the selected methods into 3 groups. The first group includes undersampling data balancing algorithms. These are SPY,[30] NearMiss,[31] TomekLinks,[32] RandomUnderSampler, OneSidedSelection,[33] and NeighbourhoodCleaningRule.[33] The second group consists of oversampling data balancing methods. These methods are SMOTE,[34] Borderline-SMOTE,[35] Clustering-based Under-sampling and Over-sampling Using Synthetic Minority Over-sampling Technique (CURE-SMOTE),[36] SMOTE-TomekLinks,[37] Automated Noise Detection Synthetic Minority Over-sampling Technique (AND-SMOTE),[38] SMOTE-D,[39] Random-SMOTE,[40] Kmean-SMOTE,[41] Gaussian-SMOTE,[24] and SMOTE-WB.[42] The third group comprises techniques from previous research for data balancing prior to machine learning. We implemented these methods on 2 standardized datasets, HNdataset and FNdataset. All experiments were conducted under identical conditions to ensure a fair comparison.

The synthetic minority oversampling technique (SMOTE), widely recognized for its ability to generate synthetic samples for the minority class by creating new data along the line connecting a minority class instance and a certain number of its same-class neighbors, has shown substantial potential in mitigating data imbalance issues. Further advancing this, the synthetic minority oversampling technique with boosting and noise detection (SMOTE-WB) represents a hybrid approach that combines SMOTE and random oversampling (ROS), incorporating additional boosting and noise detection techniques. The objective of this combination is to enhance the efficacy of synthetic sample creation, thereby improving the accuracy of classification models. The boosting technique amplifies the features of synthetic data samples by focusing on difficult-to-classify cases, while noise detection minimizes the impact of noisy data on the training process, ensuring the high quality of synthetic samples.

Table 3. The performance of the model using undersampling techniques with the FNdataset and HNdatase.

	Oversampling techniques with the FNdataset			Oversampling techniques with the HNdataset		
Method	F1	G-Mean	PR-AUC	F1	G-Mean	PR-AUC
SPY[30]	76.95%	84.21%	81.17%	81.37%	86.5%	85.62%
NearMiss[31]	19.14%	47.23%	15.33%	23.18%	58.56%	23.61%
TomekLinks[32]	79.14%	86.00%	83.29%	83.98%	88.73%	87.90%
RandomUnderSampler	64.15%	77.04%	62.85%	68.35%	78.61%	67.76%
OneSidedSelection[33]	79.21%	86.32%	83.25%	83.93%	89.11%	87.89%
NeighbourhoodCleaningRule[33]	69.82%	79.64%	72.20%	79.79%	88.09%	81.44%

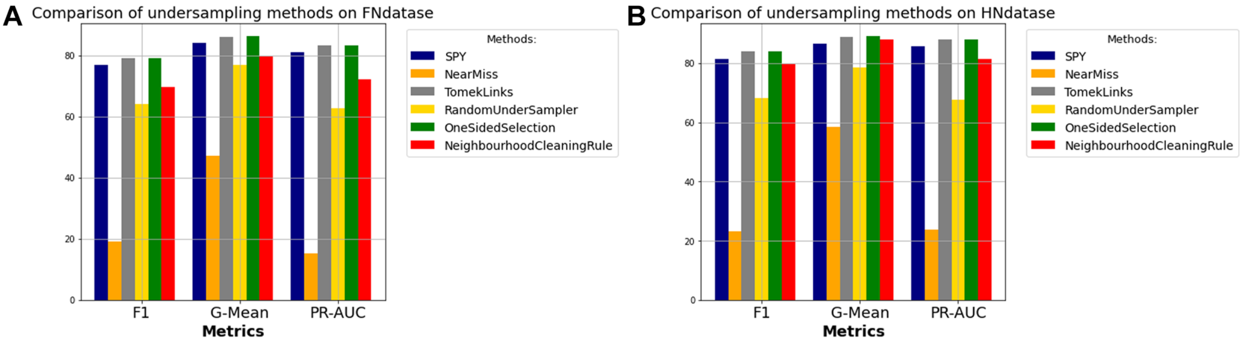

Figure 1. The performance of the model using undersampling techniques. (A) The performance of the model using undersampling techniques with the FNdatase. (B) The performance of the model using undersampling techniques with the HNdataset.

Performance analysis

Evaluation metrics. In the evaluation of machine learning models, a variety of parameters are proposed to assess performance accurately. Selecting appropriate parameters that align with each model and dataset characteristic is crucial. In scenarios involving severely imbalanced datasets, sensitivity (SE) and specificity (SP) are frequently used metrics, see equations (27) and (28) in Appendix 1. Kubat and Matwin[33] introduced the geometric mean (G-Mean) to assess machine learning models on imbalanced data (equation (29)). In addition, metrics such as accuracy (ACC) (equation (30)), recall (REC) (equation (31)), precision (PRE) (equation (32)), F1-score (equation (33)), area under the precision-recall curve (AUPR), Matthews correlation coefficient (MCC) (equation (34)), area under the curve (AUC), and precision-recall area under curve (PR-AUC) are used to evaluate and compare the effectiveness of methodologies against recent research. In this section, we evaluate the performance of our method relative to 7 prominent studies previously introduced, each designed to predict drug-disease associations (DDAs) using heterogeneous networks. Here is a brief overview of each methodology.

All experiments were conducted on a system running Microsoft Windows 11 Pro (Build 22631) with an Intel Core i5-12400 processor and 16 GB of DDR4 RAM. The experiments used Python 3.11.5 and Scikit-learn 1.3.0 for machine learning model development and evaluation. The smote_variants library[43] was employed to implement various oversampling techniques for handling imbalanced datasets. All oversampling methods were applied using the library's default parameter settings, as these configurations are well-documented and have been validated in prior studies. This choice ensures consistency and reproducibility across experiments while focusing on the evaluation of the proposed method.

Performance of high negative samples and oversampling. First, we explored the undersampling technique, applying it to both the HN dataset and the FN dataset. To enhance the reliability of the performance outcomes, the 5-fold cross-validation framework used in this study is designed to ensure an objective evaluation of the model's performance. By treating all drug-disease relationships in the test dataset as unknown during training, we ensured complete independence between training and testing processes. This practice mimics real-world scenarios, where the model is expected to predict novel drug-disease interactions without prior knowledge. We used 3 model evaluation metrics: F1, G-Mean, and PR-AUC. The results for each metric, corresponding to each undersampling method on each dataset, are detailed in Table 3 and illustrated in Figure 1.

Overall, superior results were observed with the HNdataset. Specifically, the TomekLinks method on the HNdataset produced an F1 score (equation (33)) of 83.98% and a PR-AUC of 87.90%, while the OneSidedSelection method on the HNdataset achieved a G-Mean of 89.11%.

Figure 1 displays the area beneath the PRE and REC curves for both datasets, with Figure 1A for the FNdataset

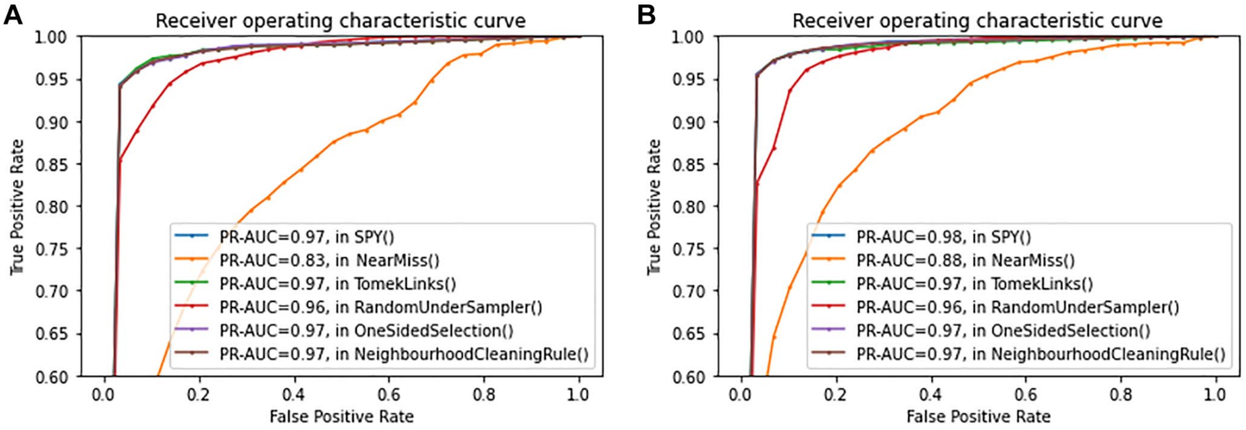

Figure 2. ROC curve with FNdataset and HNdataset of undersampling technique. (A) ROC curve with FNdataset. (B) ROC curve with HNdataset.

Table 4. Performance of the model using oversampling techniques with the FNdataset and HNdataset.

	Oversampling techniques with the FNdataset			Oversampling techniques with the HNdataset		
Method	F1	G-Mean	PR-AUC	F1	G-Mean	PR-AUC
SMOTE[34]	73.08%	79.09%	75.31%	79.85%	85.86%	83.95%
Boderline-SMOTE[35]	72.06%	78.54%	73.45%	82.17%	87.61%	83.51%
CURE-SMOTE[36]	79.20%	85.56%	83.20%	84.53%	88.58%	88.36%
SMOTE-TomekLinks[37]	72.92%	79.07%	75.16%	79.88%	85.28%	83.93%
AND-SMOTE[38]	70.57%	77.08%	71.87%	77.94%	85.65%	80.63%
SMOTE-D[39]	79.11%	85.89%	82.88%	84.24%	88.48%	88.31%
Random-SMOTE[40]	73.84%	80.61%	77.28%	81.68%	88.42%	85.41%
Kmean-SMOTE[41]	79.40%	86.86%	83.46%	84.26%	88.51%	88.09%
SMOTEWB[42]	77.57%	85.27%	81.52%	83.71%	89.70%	87.67%
Gaussian-SMOTE[24]	79.36%	86.67%	83.27%	**85.09%**	**89.80%**	**88.39%**

The best scores are printed in bold.

and Figure 1B for the HNdataset. The PR-AUC values from these curves underscore the model's high precision and recall levels for most methods applied to the HNdataset, indicating a robust capability to accurately identify positive cases across a broad range of thresholds, especially within the context of the HNdataset.

Figure 2 displays the area beneath the Precision-Recall (PR) curve for both datasets, with Figure 2A for the FNdataset and Figure 2B for the HNdataset. The PR-AUC values from these curves underscore the model's high precision and recall levels for most methods applied to the HNdataset, indicating a robust capability to accurately identify positive cases across a broad range of thresholds, especially within the context of the HNdataset.

Following that, we examined the HNdataset and FNdataset using an oversampling approach derived from a SMOTE variant. The results of these experiments are compiled in Table 4 and depicted in Figure 3. It is clear that all evaluated performance metrics for the HNdataset significantly surpassed those for the Fndataset.

In particular, the CURE-SMOTE method on the HNdataset achieved notable results, with an F1 score of 84.53% and a PR-AUC of 88.36%. Concurrently, the SMOTEWB[42] technique on the same dataset produced the highest G-Mean of 89.70% and a competitive PR-AUC of 87.67%. As shown in Tables 3 and 4, the HNdataset consistently outperformed the FNdataset across all metrics. These results highlight the effectiveness of oversampling techniques, particularly SMOTEWB,[40] in handling imbalanced datasets and improving model performance.

Figure 4 illustrates the area under the Precision-Recall (PR) curve for both datasets, with Figure 4A representing the FNdataset and Figure 4B depicting the HNdataset within the oversampling approach derived from a SMOTE variant. The derived PR-AUC metrics from these curves highlight the high levels of precision and recall achieved by most methods when applied to the HNdataset. This suggests a strong ability of the model to reliably identify positive instances over a diverse range of thresholds, particularly in the case of the HNdataset.

To evaluate the performance of the proposed methods, we compared 4 primary configurations: (1) Original, using the raw dataset without applying High Negative Dataset or Gaussian-SMOTE; (2) Gaussian-SMOTE, applying the

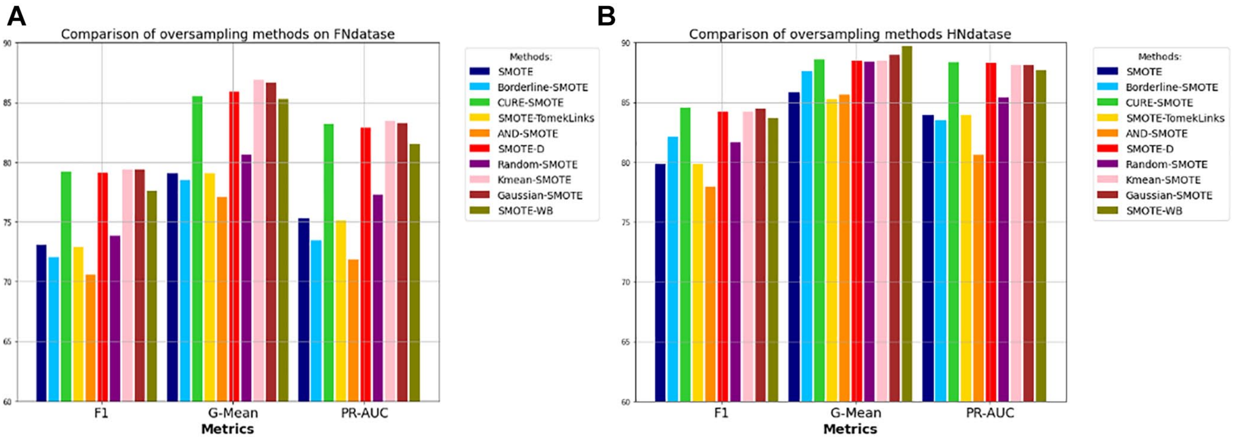

Figure 3. The performance of the variant of the SMOTE model. (A) Performance of the model using oversampling techniques with the FNdatase. (B) The performance of the model using oversampling techniques with the HNdataset.

Gaussian-SMOTE technique without using High Negative Dataset; (3) High Negative, using the High Negative Dataset without applying Gaussian-SMOTE; and (4) Our Method, which combines both High Negative Dataset and Gaussian-SMOTE. All experiments were conducted using a 5-fold cross-validation procedure to ensure consistency and fairness in evaluation. The results indicate that our method outperforms all baselines across the 3 evaluation metrics, achieving an F1-score of 85.09%, a G-mean of 89.80%, and a PR-AUC of 88.39% (Table 5).

In addition, we conducted a 2-sample t test to assess the statistical significance of the improvements. The results (Table 6) show that our method significantly outperforms Gaussian-SMOTE across all metrics with P-values $<.0002$. When compared with the High Negative Dataset, our method demonstrates statistically significant improvements with P-values $<.05$ for F1-score and PR-AUC, and approaches significance for G-mean ($P=.000313$). These findings confirm that the combination of High Negative Dataset and Gaussian-SMOTE provides substantial advantages in improving classification performance, particularly for imbalanced datasets.

Our proposed method is designed to leverage the strengths of both the High Negative Dataset and Gaussian-SMOTE. The High Negative Dataset enriches the model's ability to discriminate by providing high-quality negative samples, while Gaussian-SMOTE generates synthetic samples from the minority class, reducing data imbalance and enhancing generalization. Experimental results demonstrate that this combination not only improves performance but also yields a more stable and reliable model compared with the baseline, as evidenced by its superiority across the 3 key metrics: F1-score, G-mean, and PR-AUC.

These findings emphasize the effectiveness of oversampling methods, particularly SMOTE variants, in enhancing model performance on datasets with imbalances. Our meticulous testing process demonstrated that, across all methods, balancing the data when combined with the proposed negative sampling method consistently yields superior performance, notably enhancing the F1, G Mean, and PR-AUC metrics.

Evaluation of the proposed method against recent studies. In this section, we evaluate the performance of our method relative to 7 prominent studies previously introduced, each designed to predict drug-disease associations (DDAs) using heterogeneous networks. Here is a brief overview of each methodology.

Deep drug repositioning (deepDR)[44] employs a network-based deep learning framework to repurpose drugs in silico by integrating 10 related networks and using a multimodal deep autoencoder to learn and transform drug features into a lower-dimensional representation. A variational autoencoder encodes and decodes these features along with clinically reported drug-disease pairs to predict new applications for approved drugs. The drug-drug associations by using geometric deep learning (DDAGDL)[25] framework applies geometric deep learning to a heterogeneous information network to predict drug-drug associations, incorporating biological data and an attention mechanism to effectively manage the non-Euclidean structure of biomedical networks. Heterogeneous information network graph representation learning (HINGRL)[45] leverages a heterogeneous information network that integrates biological knowledge with drug-disease, drug-protein, and protein-disease relationships, employing graph representation learning and a Random Forest classifier to enhance drug repositioning. HNet-DNN[46] proposes a deep neural network approach using a drug-disease heterogeneous network that constructs drug-drug and disease-disease similarity networks, integrates them with existing drug-disease associations, extracts topological features, and uses them to train the DNN.

The drug repositioning approach based on weighted bilinear neural collaborative filtering (DRWBNCF)[47] employs a deep learning model that integrates various similarity networks and uses a novel weighted bilinear graph convolution technique along with a multilayer perceptron

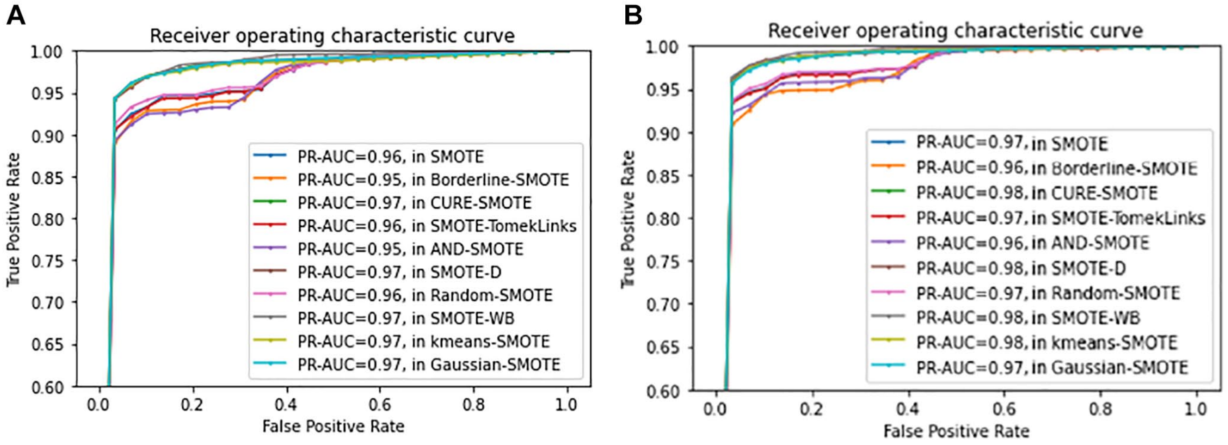

Figure 4. ROC curve with FNdataset and HNdataset of oversampling techniques. (A) ROC curve with FNdataset. (B) ROC curve with HNdataset.

Table 5. Classification performance F1-score, G-mean, and PR-AUC of different methods.

	F1	G-mean	PR-AUC
Original	52.64%	69.76%	51.67%
High negative	83.39%	87.79%	87.46%
Gaussian-SMOTE	54.13%	69.33%	52.48%
Our method	85.09%	89.80%	88.39%

optimized by specific loss functions and graph regularization to enhance drug repositioning by predicting new drug-disease relationships. Drug Repositioning based on the Heterogeneous information fusion graph convolutional network (DRHGCN)[48] uses graph convolutional networks to analyze and integrate data from drug-drug, disease-disease, and drug-disease association networks, employing a layer attention mechanism to refine features from multiple network layers. Attention-aware multi-modal fusion using a dual-graph transformer (AMDDT)[49] is based on dual-graph transformer modules, leveraging advanced graph neural networks for predicting drug-disease associations.

In this study, we consistently applied 5-fold cross-validation across all experiments, including those that combined the high negative sampling (HNS) and full negative sampling (FNS) techniques. The comparison of model performance with HNS and FNS is presented in the last 2 rows of Table 7. The results demonstrate that the HNS strategy significantly outperforms FNS across all performance metrics. Specifically, AUPR improved from 0.892 to 0.915, AUC increased from 0.959 to 0.966, PRE rose from 0.835 to 0.862, REC increased from 0.843 to 0.851, ACC improved from 0.932 to 0.938, MCC increased from 0.793 to 0.817, and the F1-score rose from 0.835 to 0.856. These substantial improvements underscore the effectiveness of the HNS strategy in leveraging informative negative samples to enhance model training quality.

Furthermore, compared with prior studies, our method consistently demonstrated superior performance, as shown in Table 7. Notable results achieved by our approach include AUPR: 0.980, AUC: 0.983, REC: 0.940, ACC: 0.946, MCC: 0.890, and F1-score: 0.935. These metrics surpass all existing approaches, providing compelling evidence of the optimized predictive capability of our model when employing the HNS strategy for drug-disease interaction prediction. We have uploaded the complete implementation, along with a detailed README file, to GitHub: https://github.com/hunglm11/BI-DD-HNSO.

These findings not only confirm the effectiveness of the HNS strategy but also highlight its potential applications in other complex predictive tasks. This approach maximizes the extraction of valuable information from potential negative samples, enhancing both the accuracy and reliability of the model. The integration of an advanced model with an efficient negative sampling strategy offers promising avenues for drug-disease interaction research, contributing to the advancement of prediction methodologies in this domain.

The exceptional performance of our method can largely be attributed to the integration of oversampling technique with a high negative approach. This innovative strategy significantly enhances the accuracy of the model by ensuring that the trained classifier does not overly fit the majority class in imbalanced datasets, which is a common challenge in medical data analysis. Moreover, this approach allows precise capturing of rare but medically significant patterns within the data, contributing to the model's high recall and precision rates. The successful application of these techniques ensures that our method not only performs well in balanced scenarios but excels even under conditions characterized by data sparsity and imbalance, a common issue in the complex landscape of drug-disease association prediction.

Case study. The relationships between drugs and diseases extracted from DrugBank have been identified with 1187 positive samples and supplemented by 530 687 unknown drug-disease pairs categorized as high-negative samples. To balance the dataset for model training, the SMOTE-WB technique was employed. The model was then trained on

Table 6. Statistical significance of improvements via 2-sample *t* test.

	F1			G-mean			PR-AUC		
	Gaussian-SMOTE	High negative	Our method	Gaussian-SMOTE	High negative	Our method	Gaussian-SMOTE	High negative	Our method
Original	1.0E-05	3.7E-01	5.0E-06	1.2E-04	7.8E-01	5.1E-05	6.0E-06	6.6E-02	6.0E-06
Gaussian-SMOTE	**x**	5.0E-06	2.0E-06	**x**	3.7E-04	1.9E-04	**x**	5.0E-06	4.0E-06
High negative	5.0E-06	**x**	3.1E-04	3.7E-04	**x**	4.4E-04	5.0E-06	**x**	1.8E-04
Our method	2.0E-06	3.1E-04	**x**	1.9E-04	4.4E-04	**x**	4.0E-06	1.8E-04	**x**

Table 7. Comparison of the performance of results against recent studies on the B-dataset.

Method*	AUPR	AUC	PRE	REC	ACC	MCC	F1
deepDR[44]	0.804	0.820	0.883	0.233	0.601	0.299	0.369
DDAGDL[25]	0.831	0.842	0.761	0.770	0.764	0.529	0.765
HINGRL[45]	0.877	0.884	0.800	0.808	0.803	0.607	0.804
Hnet-DNN[46]	0.891	0.892	0.782	0.828	0.810	0.621	0.804
DRWBNCF[47]	0.901	0.900	**0.981**	0.202	0.599	0.326	0.335
DRHGCN[48]	0.910	0.909	0.867	0.771	0.826	0.658	0.816
AMDGT[49]	0.930	0.933	0.861	0.865	0.862	0.725	0.863
Model with FNS	0.892	0.959	0.835	0.843	0.932	0.793	0.835
Our method	0.915	**0.966**	0.862	**0.851**	**0.938**	**0.817**	**0.856**

*The best scores are printed in bold.

these unknown pairs to predict potential drug-disease relationships. Rigorous validation of the results was conducted through a review of credible biomedical literature.

The top 20 predictions from the model are presented in Table 8, with 12 of these predictions being substantiated by authoritative biomedical reports. Notably, Levobunolol has been shown effective in treating various forms of glaucoma, including Primary Open Angle Glaucoma (Disease ID: 137760) and Glaucoma 1, Primary Open Angle (Disease ID: 601682). Gliclazide has demonstrated efficacy against several types of Maturity-Onset Diabetes of the Young (MODY), such as MODY3 (Disease ID: 600496), MODY1 (Disease ID: 125850), and MODY2 (Disease ID: 125851). In addition, the relationships of Triamcinolone and Prednisone with autoimmune and inflammatory disorders like Multiple Sclerosis (Disease ID: 126200) and Otitis Media (Disease ID: 166760), as well as Estradiol's connection to Hereditary Prostate Cancer type 1 (Disease ID: 601518), highlight the intricate interactions between drug mechanisms and disease pathologies. The use of Betamethasone in treating Hydrocortisone in Sarcoidosis (Disease ID: 181000), Cimetidine in Helicobacter Pylori infections (Disease ID: 600263), and Daunorubicin in Classic Hodgkin Lymphoma (Disease ID: 236000) further illustrates the significant role these drugs play in specific disease management strategies. Even though some predicted drug-disease associations lack direct documentary evidence, these findings pave new pathways for potential clinical trials and research, potentially reducing both the time and cost involved in drug development processes.

These findings not only validate the predictive model but also lay the groundwork for further clinical research and the development of personalized medical treatments, thereby enhancing the accuracy and efficiency of disease management protocols.

Conclusions

In this study, we proposed a novel approach by combining data balancing techniques with the high-confidence negative sample selection method to predict the relationship between drugs and diseases. Our findings demonstrated promising results, indicating the efficacy of this approach in enhancing the reliability of predictive models. First, we employed Bayesian theory to analyze the relationships between drugs and diseases, constructing a heterogeneous drug-protein-disease network. From this, we extracted richly informative drug-disease feature vectors. Second, we demonstrated that our method of selecting high-confidence negative samples effectively eliminated unreliable negative instances, contributing significantly to the overall improvement of model performance. This enhancement not only increased the accuracy of predictions but also fostered trust in the model's outputs. Furthermore, by recommending certain drugs for diseases, some of which have been scientifically validated through clinical trials and are currently employed in treatment regimens, we have bolstered the credibility of our model's predictions.

Looking ahead, we aim to continue refining our method for selecting reliable samples and enhancing data balancing techniques to further improve the efficiency of our model. This ongoing pursuit of refinement will ensure that our predictive model remains at the forefront of precision medicine, offering valuable insights into drug-disease relationships for clinical decision-making and therapeutic advancements.

Table 8. Top-20 candidate drugs for various diseases.

Rank	Drug ID	Drug name	Disease ID	Disease name	Prob*	Literature validation**
1	DB01120	Gliclazide	600496	Maturity-Onset Diabetes Of The Young, Type 3; Mody3	0.982	Habeb et al,[50] Spiliotis et al[51]
2	DB00783	Estradiol	192000	Uterine Anomalies	0.973	NA
3	DB01120	Gliclazide	125850	Maturity-Onset Diabetes Of The Young, Type 1; Mody1	0.967	Habeb et al,[50] Spiliotis et al[51]
4	DB01120	Gliclazide	125851	Maturity-Onset Diabetes Of The Young, Type 2; Mody2	0.966	Habeb et al,[50] Spiliotis et al[51]
5	DB01210	Levobunolol	137760	Glaucoma, Primary Open Angle; Poag	0.948	Sorensen et al[52]
6	DB01210	Levobunolol	601682	Glaucoma 1, Primary Open Angle, C; Glc1C	0.946	Sorensen et al[52]
7	DB00232	Methyclothiazide	600351	Enteropathy, Familial, With Villous Edema And Immunoglobulin G2 Deficiency	0.944	NA
8	DB00620	Triamcinolone	126200	Multiple Sclerosis, Susceptibility To; Ms	0.943	Lukas et al,[53] Abu-Mugheisib et al[54]
9	DB00712	Flurbiprofen	133690	Exostoses With Anetodermia And Brachydactyly, Type E	0.941	NA
10	DB00590	Doxazosin	157950	Permanent Molars, Secondary Retention Of	0.940	NA
11	DB01070	Dihydrotachysterol	259660	Malignant Hyperthermia, Susceptibility To, 3	0.930	NA
12	DB00635	Prednisone	166760	Otitis Media, Susceptibility To; Oms	0.934	Ranakusuma et al[55]
13	DB00783	Estradiol	601518	Prostate Cancer, Hereditary, 1; Hpc1	0.926	Ockrim et al[56]
14	DB00443	Betamethasone	188030	Immune Thrombocytopenia; Itp	0.926	NA
15	DB00741	Hydrocortisone	181000	Sarcoidosis, Susceptibility To, 1; Ss1	0.925	Sullivan et al[57]
16	DB00481	Raloxifene	215470	Boucher-Neuhauser Syndrome; Bnhs	0.925	NA
17	DB01013	Clobetasol propionate	233810	Growth Retardation, Small And Puffy Hands And Feet, And Eczema	0.925	NA
18	DB00501	Cimetidine	600263	Helicobacter Pylori Infection, Susceptibility To	0.919	Higuchi et al[58]
19	DB00694	Daunorubicin	236000	Lymphoma, Hodgkin, Classic; Chl	0.917	Richardson et al[59]
20	DB00136	Calcitriol	241519	Hypophosphatemia, Renal, with Intracerebral Calcifications	0.913	NA

*The Prob represents the predicted probability of each drug's effectiveness for diseases in the test dataset, based on equation (21). **NA stands for non-availability of literature evidence for drug-disease pairs.

Acknowledgments

We acknowledge support from the Electric Power University and Academy of Policy and Development.

ORCID iDs

Nam Anh Dao https://orcid.org/0000-0002-0536-8686
Xuan Tho Dang https://orcid.org/0000-0002-7654-5942

Statements and Declarations

Author Contributions

Manh Hung Le: Methodology; Software; Data curation; Writing – Results and discussion.
Nam Anh Dao: Conceptualization of this study; Methodology.
Xuan Tho Dang: Data curation; Methodology; Writing – introduction; Related woks.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. *Brief Bioinform*. 2016;17:2-12. doi:10.1093/bib/bbv020
2. Nelson BS, Kremer DM, Lyssiotis CA. New tricks for an old drug. *Nat Chem Biol*. 2018;14:990-991. doi:10.1038/s41589-018-0137-x
3. Tho Dang X, Hung Le M, Anh Dao N. Drug repositioning for drug disease association in meta-paths. In: Phuong NH, Kreinovich V eds. *Deep Learning and Other Soft Computing Techniques. Studies in Computational Intelligence* (Vol. 1097). Springer; 2022. doi:10.1007/978-3-031-29447-1_4
4. Kun L, Jia W, Bo D, et al. Towards a better model with dual transformer for drug response prediction. arXiv. 2024. doi:10.48550/arXiv.2210.17401v2
5. Gu Y, Zheng S, Zhang B, Kang H, Jiang R, Li J. Deep multiple instance learning on heterogeneous graph for drug-disease

association prediction. *Comput Biol Med*. 2025;184:109403. doi:10.1016/j.compbiomed.2024.109403

6. You Y, Lai X, Pan Y, et al. Artificial intelligence in cancer target identification and drug discovery. *Sig Transduct Target Ther*. 2022;7:1. doi:10.1038/s41392-022-00994-0
7. Yu Z, Huang F, Zhao X, Xiao W, Zhang W. Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform*. 2021;22:bbaa243. doi:10.1093/bib/bbaa243
8. Thai TV, Bui DH, Dang XT, Nguyen TP, et al. A new computational method based on heterogeneous network for predicting MicroRNA-disease associations. In: Kreinovich V, Hoang Phuong N, eds. *Soft Computing for Biomedical Applications and Related Topics*. 2020:205-219. doi:10.1007/978-3-030-49536-7_18
9. Li Z, Huang Q, Chen Y, Wang J, et al. Identification of drug-disease associations using information of molecular structures and clinical symptoms via deep convolutional neural network. *Front Chem*. 2020;8:924. doi:10.3389/fchem.2019.00924
10. Dao NA, Bui DH, Pham QH, Dang XT. Feature analysis for imbalanced learning. *J Adv Comput Intell Intell Inform*. 2020;24:648-655. doi:10.20965/jaciii.2020.p0648
11. Varothai S, Bergfeld WF. Androgenetic alopecia: an evidence-based treatment update. *Am J Clin Dermatol*. 2014;15:217-230. doi:10.1007/s40257-014-0077-5
12. Novac N. Challenges and opportunities of drug repositioning. *Trends Pharmacol Sci*. 2013;34:267-272. doi:10.1016/j.tips.2013.03.004
13. Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS ONE*. 2013;8:e78518. doi:10.1371/journal.pone.0078518
14. Khalid Z, Sezerman OU. Computational drug repurposing to predict approved and novel drug-disease associations. *J Mol Graph Model*. 2018;85:91-96. doi:10.1016/j.jmgm.2018.08.005
15. Zhang W, Yue X, Lin W, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics*. 2018;19:1-12. doi:10.1186/s12859-018-2220-4
16. Yang J, Li Z, Fan X, Cheng Y. Drug-disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. *J Chem Inf Model*. 2014;54:2562-2569. doi:10.1021/ci500340n
17. Liu H, Song Y, Guan J, Luo L, Zhuang Z. Inferring new indications for approved drugs via random walk on drug-disease heterogeneous networks. *BMC Bioinformatics*. 2016;17:269-277. doi:10.1186/s12859-016-1336-7
18. Zhang W, Yue X, Huang F, Liu R, Chen Y, Ruan C. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods*. 2018;145:51-59. doi:10.1016/j.ymeth.2018.06.001
19. Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*. 2020;36:1241-1251. doi:10.1093/bioinformatics/btz718
20. Wu G, Liu J, Yue X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinformatics*. 2020;20:134. doi:10.1186/s12859-019-2644-5
21. Liu J, Zuo Z, Wu G. Link prediction only with interaction data and its application on drug repositioning. *IEEE Trans Nanobioscience*. 2020;19:547-555. doi:10.1109/TNB.2020.2990291
22. Barber D. *Bayesian Reasoning and Machine Learning*. Cambridge University Press; 2010. doi:10.1017/CBO9780511814774
23. Gonzalez MW, Kann MG. Chapter 4: protein interactions and disease. *PLoS Comput Biol*. 2012;8:e1002819. doi:10.1371/journal.pcbi.1002819
24. Hansoo L, Jonggeun K, Sungshin K. Gaussian-based SMOTE algorithm for solving skewed class distributions. *Int J Fuzzy Logic Intell Syst*. 2017;17:229-234. doi:10.5391/IJFIS.2017.17.4.229
25. Zhao BW, Su XR, Hu PW, Ma YP, Zhou X, Hu L. A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Brief Bioinform*. 2022;23:bbac384. doi:10.1093/bib/bbac384
26. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33:d514-d517. doi:10.1093/nar/gki033
27. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46:D1074-D1082. doi:10.1093/nar/gkx1037
28. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*. 2011;7:496. doi:10.1038/msb.2011.26
29. Le MH, Dao NA, Dang XT. High potential negative sampling for drug-disease association prediction. In: Phuong NH, Chau NT, Kreinovich V, eds. *Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications. Studies in Systems, Decision and Control* (Vol. 543). Springer; 2024:120-134. doi:10.1007/978-3-031-63929-6_7
30. Dang XT, Tran DH, Hirose O, Satou K. SPY: a novel resampling method for improving classification performance in imbalanced data. Paper presented at: Seventh International Conference on Knowledge and Systems Engineering (KSE); October 8-10, 2015; Ho Chi Minh City, Vietnam. doi:10.1109/KSE.2015.24
31. Zhang JP, Mani I. KNN approach to unbalanced data distributions: a case study involving information extraction. Paper presented at: International Conference on Machine Learning (ICML); June 21-24, 2003; Washington, DC.
32. Tomek I. Two modifications of CNN. *IEEE Trans Syst Man Cybern*. 1976;6:769-772. doi:10.1109/TSMC.1976.4309452
33. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. Paper presented at: International Conference on Machine Learning (ICML); July 8-12, 1997; Nashville, TN.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953
35. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Adv Intell Comput*. 2005;3644:878-887. doi:10.1007/11538059_91
36. Ma L, Fan S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics*. 2017;18:169. doi:10.1186/s12859-017-1578-z
37. Batista G, Bazzan B, Monard M. Balancing training data for automated annotation of keywords: a case study. In: *WOB*. 2003. http://dblp.uni-trier.de/db/conf/wob/wob2003.html#BatistaBM03

38. Yun J, Ha J, Lee JS. Automatic determination of neighborhood size in SMOTE. Paper presented at: 10th International Conference on Ubiquitous Information Management and Communication; January 21-23, 2016; Da Nang, Vietnam. doi:10.1145/2857546.2857648
39. Torres FR, Carrasco-Ochoa JA, Martinez-Trinidad JF. SMOTE-D: a deterministic version of SMOTE. In: Martínez-Trinidad JF, Carrasco-Ochoa JA, Ayala-Ramirez V, eds. *Pattern Recognition. Lecture Notes in Computer Science* (Vol. 9703). Springer; 2016:134-143. doi:10.1007/978-3-319-39393-3_18
40. Dong Y, Wang XA. New over-sampling approach: random-SMOTE for learning from imbalanced data sets. In: Wang L, ed. *Knowledge Science, Engineering and Management. Lecture Notes in Computer Science* (Vol. 7091). Springer; 2011:471-482. doi:10.1007/978-3-642-25975-3_30
41. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inform Sci*. 2018;467:1-20. doi:10.1016/j.ins.2018.07.040
42. Saglam F, Cengiz MA. A novel SMOTE-based resampling technique through noise detection and the boosting procedure. *Expert Syst Appl*. 2022;192:117023. doi:10.1016/j.eswa.2022.117023
43. Kovács G. Smote-variants: a Python implementation of 85 minority oversampling techniques. *Neurocomputing*. 2019;366:352-354. doi:10.1016/j.neucom.2019.06.100
44. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. DeepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*. 2019;35:5191-5198. doi:10.1093/bioinformatics/btz418
45. Zhao BW, Hu L, You ZH, Wang L, Su XR. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Brief Bioinform*. 2021;23:bbab515. doi:10.1093/bib/bbab515
46. Liu H, Zhang W, Song Y, Deng L, Zhou S. HNet-DNN: inferring new drug-disease associations with deep neural network based on heterogeneous network features. *J Chem Inf Model*. 2020;60:2367-2376. doi:10.1021/acs.jcim.9b01008
47. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform*. 2022;23:bbab581. doi:10.1093/bib/bbab581
48. Cai L, Lu C, Xu J, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform*. 2021;22:bbab319. doi:10.1093/bib/bbab319
49. Liu J, Guan S, Zou Q, Wu H, Tiwari P, Ding Y. AMDGT: attention-aware multi-modal fusion using a dual graph transformer for drug-disease associations prediction. *Brief Bioinform*. 2022;23:bbab515. doi:10.1016/j.knosys.2023.111329
50. Habeb AM, George ET, Mathew V, Hattersley AL. Response to oral gliclazide in a pre-pubertal child with hepatic nuclear factor-1 alpha maturity onset diabetes of the young. *Ann Saudi Med*. 2011;31:190-193. doi:10.4103/0256-4947.75590
51. Spiliotis II, Anguelova L, Kavvoura F, Owen K. OR34-06 gliclazide restores appropriate glucagon suppression during OGTT in MODY3 & MODY1 patients: results of the “Glucagon in MODY” study. *J Endocr Soc*. 2023;7:bvad1141045. doi:10.1210/jendso/bvad114.1045
52. Sorensen SJ, Abel SR. Comparison of the ocular beta-blockers. *Ann Pharmacother*. 1996;30:43-54. doi:10.1177/106002809603000109
53. Lukas C, Bellenberg B, Hahn HK, et al. Benefit of repetitive intrathecal triamcinolone acetonide therapy in predominantly spinal multiple sclerosis: prediction by upper spinal cord atrophy. *Ther Adv Neurol Disord*. 2009;2:42-49. doi:10.1177/1756285609343480
54. Abu-Mugheisib M, Benecke R, Zettl UK. Repeated intrathecal triamcinolone acetonide administration in progressive multiple sclerosis: a review. *Mult Scler Int*. 2011;2011:219049. doi:10.1155/2011/219049
55. Ranakusuma RW, McCullough AR, Safitri ED, et al. Oral prednisolone for acute otitis media in children: a pilot, pragmatic, randomised, open-label, controlled study (OPAL study). *Pilot Feasibility Stud*. 2020;6:121. doi:10.1186/s40814-020-00671-5
56. Ockrim JL, Lalani EN, Laniado ME, Carter SS, Abel PD. Transdermal estradiol therapy for advanced prostate cancer-forward to the past? *J Urol*. 2003;169:1735-1737. doi:10.1097/01.ju.0000061024.75334.40
57. Sullivan RD, Mayock RL, Jones RJ. Local injection of hydrocortisone and cortisone into skin lesions of sarcoidosis. *JAMA*. 1953;152:308-312. doi:10.1001/jama.1953.03690040012005
58. Higuchi K, Tanigawa T, Hamaguchi M, et al. Comparison of the effects of rebamipide with those of cimetidine on chronic gastritis associated with Helicobacter pylori in Mongolian gerbils. *Aliment Pharmacol Ther*. 2003;18:1-7. doi:10.1046/j.1365-2036.2003.018s101.x
59. Richardson DS, Kelsey SM, Johnson SA, Tighe M, Cavenagh JD, Newland AC. Early evaluation of liposomal daunorubicin (DaunoXome, Nexstar) in the treatment of relapsed and refractory lymphoma. *Invest New Drugs*. 1997;15:247-253. doi:10.1023/a:1005879219554

Appendix 1

Here, $TP, FP, TN, and\ FN$ represent the number of true positives, false positives, true negatives, and false negatives, respectively. The formula for calculating the efficiency metric of the model is expressed as follows:

$$SE = TP / (TP + FN), \tag{27}$$

$$SP = TN / (FP + TN), \tag{28}$$

$$G - Mean = \sqrt{SP \times SE}, \tag{29}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{30}$$

$$REC = \frac{TP}{TP + PN}, \tag{31}$$

$$PRE = \frac{TP}{TP + FP}, \tag{32}$$

$$F1 - score = \frac{2 \times REC \times PRE}{REC + PRE}, \tag{33}$$

$$MCC = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}. \tag{34}$$