SVM-SMOTE Kullanarak İlaç-Hedef Etkileşimi Tahminini İyileştirme: Dengesiz Veri Setleri İçin Bir Çözüm

***

Improving Drug-Target Interaction Prediction Using SVM-SMOTE: A Solution for Imbalanced Dataset

Sara NAGHIB ZADEH[1][2]
Zümrüt ECEVİT SATI[3]
Ali GHANBARİ SORKHI[4]

DOI:10.33461/uybisbbd.1661593

Makale Bilgileri

Makale Türü:
Araştırma Makalesi

Geliş Tarihi:
20.03.2025

Kabul Tarihi:
27.05.2025

Öz

İlaç-hedef etkileşimi (DTI) tahmini, ilaç keşfi sürecinin kritik bir aşamasıdır çünkü deneysel yöntemler genellikle zaman alıcı ve maliyetlidir. Bu görev için makine öğrenimi teknikleri etkili alternatifler olarak ortaya çıkmıştır. Ancak, DTI veri kümeleri genellikle ciddi bir sınıf dengesizliği sorunu yaşar; gerçek etkileşimlerin sayısı negatif örneklerden önemli ölçüde azdır ve bu durum model eğitimi için ciddi bir zorluk oluşturur.Bu çalışma, DTI tahmini için etkili bir çerçeve önermektedir. Model, protein özelliklerini çıkarmak için amino asit kompozisyonu (AAC) ve dipeptit kompozisyonu (DPC) yöntemlerini kullanırken, ilaç özelliklerini temsil etmek için FP2 moleküler parmak izlerinden yararlanır. Sınıf dengesizliği sorununu ele almak amacıyla, destek vektör makineleri (SVM) tabanlı sentetik azınlık çoğaltma yöntemi olan SVM-SMOTE tekniği uygulanmıştır. Modelin eğitimi için Lineer Destek Vektör Makineleri (LSVM) algoritması kullanılmıştır. Önerilen model, Enzyme, GPCR, Ion Channel ve Nuclear Receptor gibi standart veri kümeleri kullanılarak mevcut ileri düzey yöntemlerle karşılaştırılmış ve üstün performans sergilediği görülmüştür. Model tasarımının çeşitli aşamalarında geniş kapsamlı deneyler gerçekleştirilmiş ve AUC, doğruluk, F1 skoru ve hatırlama (recall) gibi değerlendirme metrikleri kullanılarak önerilen yaklaşımın etkinliği doğrulanmıştır.

Anahtar Kelimeler: *İlaç hedef etkileşimi, Özellik çıkarımı, Veri dengeleme, SVM_SMOTE, Doğrusal SVM.*

Abstract

Article Info

Paper Type:
Research Paper

Received:
20.03.2025

Accepted:
27.05.2025

Drug–target interaction (DTI) prediction is a critical step in the drug discovery process, as experimental methods are often time-consuming and expensive. Machine learning techniques have emerged as effective alternatives for this task. However, DTI datasets commonly suffer from severe class imbalance, where the number of true interactions is significantly lower than negative ones—posing a serious challenge for model training. This study proposes an effective framework for DTI prediction. The model utilizes amino acid composition (AAC) and dipeptide composition (DPC) methods to extract protein features, while FP2 molecular fingerprints are used to represent drug features. To address the class imbalance problem, the SVM-SMOTE technique—an SVM-based synthetic minority oversampling method—is employed. For model training, a Linear Support Vector Machine (LSVM) algorithm is used. The proposed model was evaluated against several state-of-the-art methods using benchmark datasets, including Enzyme, GPCR, Ion Channel, and Nuclear Receptor. The results demonstrate that the proposed framework achieves superior performance. Extensive experiments were conducted at various stages of model design, using evaluation metrics such as AUC, accuracy, F1-score, and recall, all of which confirm the effectiveness of the proposed approach.

Keywords: *Drug target interaction, Feature extraction, Data balancing, SVM_SMOTE, Linear SVM.*

Atıf/ to Cite (APA): Naghib Zadeh S., Ecevit Satı Z., & Ghanbari Sorkhi A. (2025). Improving Drug-Target Interaction Prediction Using SVM-SMOTE: A Solution for Imbalanced Dataset. International Journal of Management Information Systems and Computer Science, 9(1), 10-28. DOI: 10.33461/uybisbbd.1661593

[1] Istanbul University, Institute of Graduate Studies in Sciences, Department of Informatics, Türkiye.

[2] Halic University, Vocational School, Istanbul, Department of Computer Programming, Türkiye

[3] Istanbul University, Faculty of Political Sciences, Department of Business Administration, Türkiye

[4] Department of Electrical and Computer Engineering, University of Science and Technology of Mazandaran, Behshahr, Iran

1. INTRODUCTION

The process of drug discovery is fundamental to advancements in both the pharmaceutical and medical fields. A fundamental part of this process is the prediction of drug-target interactions (DTIs), which is critical for the discovery of potential drug candidates. In DTI studies, drugs are typically chemical compounds, whereas targets refer to specific proteins that interact with them. Accurate DTI prediction not only facilitates drug discovery but also contributes to the efficiency and cost-effectiveness of pharmaceutical research(Gao et al., 2018).

Given the critical importance of uncovering drug-target interactions, computational prediction methods—such as molecular docking and machine learning—have garnered increasing attention for their ability to identify novel drug-target pairs with remarkable accuracy (Gao et al., 2018) .These approaches encompass ligand-based, target-based, network-based, and machine learning strategies, each distinguished by unique advantages and inherent limitations. For instance, ligand-based methods offer speed and efficiency but are constrained in their ability to generalize; target-based approaches rely heavily on high-quality biological data; while network-based techniques provide a holistic view yet demand complex interpretation(Ikechukwu & Kumar, 2023 ; Moesgaard, 2024;Vlasiou, 2024 ; S. Gao et al., 2022) . Ultimately, machine learning—particularly through deep neural networks and transformer architectures—excels at deciphering vast and intricate datasets to uncover rare interactions, albeit at the cost of substantial computational power and extensive data requirements (L. Wang et al., 2023).

Machine learning approaches for predicting drug-target interactions (DTI) are broadly categorized into two main types: similarity-based and feature-based methods. Similarity-based approaches detect interaction patterns by evaluating various similarity criteria between drugs and analogous proteins (An & Yu, 2021; Bagherian, Kim, et al., 2021; Sorkhi et al., 2021). In contrast, feature-based methods frame the problem as a binary classification task, extracting meaningful features from drug and protein data and employing algorithms such as support vector machines, random forests, XGBoost, and deep neural networks to make predictions (Mousavian et al., 2016; Shi et al., 2019a; T. Chen & Guestrin, 2016; Hu et al., 2019). Due to their capacity for rapid processing and comprehensive analysis of large datasets, these techniques enable extensive screening of drug candidates and streamline the discovery of novel interactions, establishing themselves as a vital pillar in computer-aided drug discovery (El-Behery et al., 2022; K. Y. Gao et al., 2018).

The field of pharmacology data analysis, particularly in the detection of drug-target interactions, has undergone a significant transformation with advances in machine learning (ML), opening new avenues for improvement and development in this domain (Padhi et al., 2023). One of the fundamental challenges in computational biology is the imbalance between positive (interactive) and negative (non-interactive) classes, where the number of interactive drug-target pairs is significantly lower than that of non-interactive pairs. This imbalance leads to bias in machine learning models. A major challenge in drug-target interaction prediction is managing this data imbalance. This issue often causes models to favor classifying pairs as non-interactive, as they are more abundant, making the accurate identification of interactive pairs more difficult (Redkar et al., 2020).

The objectives of this study focus on addressing the challenges associated with drug-target interaction (DTI) prediction, with an emphasis on data imbalance and increasing the accuracy of machine learning models. One of the fundamental issues in this field is the unequal distribution of positive and negative interactions, which leads to a bias in machine learning models toward the majority class (non-interactive pairs). This issue can reduce the accuracy of real interaction predictions and pose challenges to the drug discovery process. The key objectives of this study are as follows:

- Emphasizing the importance of DTI prediction before drug development to reduce costs and enhance efficiency in the drug discovery process.

- Enhancing the discriminatory power of predictive models by extracting effective features from drugs and proteins, leading to improved accuracy in identifying real interactions.
- Balancing imbalanced data in DTI prediction using advanced techniques, particularly SVM-SMOTE, to reduce bias in machine learning models.
- Increasing the speed and accuracy of drug-target interaction predictions through an efficient computational framework that facilitates the identification of novel interactions.

By presenting a comprehensive and effective approach to DTI prediction, this study aims to contribute to the improvement of the drug discovery process and the development of more efficient methods in biomedical research.

The organization of this paper is as follows: Section 2 provides an overview of previous studies and different approaches to predict drug-target interactions. Section 3 describes the proposed method, which includes the steps of feature extraction for drugs and proteins, data preprocessing, class balancing using SVM-SMOTE, and training a linear SVM model. In Section 4, the proposed model's performance is assessed using standard evaluation metrics and compared with various machine learning algorithms and other approaches. Finally, Section 5 presents the conclusion and suggestions for future research.

2. RELATED WORKS

Drug-target interaction (DTI) prediction plays a key role in drug development. It can reduce experimental trials, accelerate drug discovery, and lower costs(Abbasi et al., 2020). Machine learning (ML) and deep learning (DL), which have been successful in various fields, are also used in bioinformatics for analyzing genetic and pharmaceutical data(Bian et al., 2023). Machine learning is generally categorized into supervised and unsupervised learning (Lo et al., 2018). In supervised learning, models utilize labeled data for training and prediction. Some of the most widely used algorithms in this domain include Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN), all of which are extensively employed in drug discovery, particularly for predicting drug-target interactions (Lo et al., 2018; Mitchell B.O., 2014).In contrast, unsupervised machine learning algorithms, such as k-Means clustering and hierarchical clustering, can identify hidden patterns in data without requiring labeled datasets, playing a significant role in DTI prediction. Due to their high flexibility and ability to detect previously unknown clusters, these methods have gained considerable attention as effective approaches in this field (Lo et al., 2018).

Machine learning techniques have gained importance in the pharmaceutical industry due to their ability to accelerate the analysis of large datasets and have now become the primary technique for predicting drug-target interactions(Charoenkwan et al., 2021a).In this context, Faulon et al. combined chemoinformatics data with SVM to predict drug-target interactions without requiring explicit binding information, successfully identifying unknown interactions at the genomic scale using signature kernels (Faulon et al., 2008).Wang et al. employed a random forest model to predict drug-target interactions, utilizing features derived from molecular vibrations and selected through the Boruta algorithm. However, the model's performance suffered due to its sensitivity to noise and the limited effectiveness of manually engineered features (X. rui Wang et al., 2021).Ezzat et al. proposed EnsemDT, a hybrid learning model for DTI prediction that integrates multiple decision tree classifiers. Drug features were extracted from SMILES and target features via PROFEAT, enabling accurate prediction using combined feature vectors (Ezzat et al., 2019).

Deep learning has proven highly effective in predicting drug-target interactions (DTIs) by extracting complex features from large-scale biomedical data. Models like CNN and MLP have been widely applied for this purpose (Azlim Khan & Ahamed Hassain Malim, 2023a).For instance, Li et al. introduced DeepConv-DTI, which uses 1D-CNN to extract protein sequence features and processes drug representations via ECFP and fully connected layers(I. Lee et al., 2019). Expanding on the concept of feature interaction, Huang et al. proposed MolTrans, which leverages Transformers to encode SMILES and protein sequences, generating 2D interaction maps subsequently processed by CNNs and FCNs (K. Huang et al., 2021).However, most of these models either ignore the explicit interaction context between drug-target pairs or rely heavily on large, balanced datasets, which are rare in real-world scenarios.

In the healthcare domain, data imbalance—often due to the relatively small number of positive cases (e.g., infected individuals) compared to negative ones—poses a major challenge to the effectiveness of machine learning models. To address this, resampling techniques have been widely employed to improve predictive accuracy. For instance, Latief et al. tackled the class imbalance issue in lung cancer diagnosis by combining SMOTE and ENN, which significantly enhanced classification performance when used with the Random Forest model (Latief et al., 2024). Similarly, Huang et al. focused on breast cancer prediction by integrating feature selection methods—Information Gain (IG) and Genetic Algorithm (GA)—with the SMOTE oversampling technique, leading to notable improvements in model accuracy on imbalanced datasets (M. W. Huang et al., 2021).In another study, to enhance the early detection of heart disease, Akkaya et alemployed a combination of the SMOTE-Tomek Links technique to address class imbalance and the XGBoost and k-NN algorithms for model training, achieving notable results(Akkaya et al., 2022).In this context, drug-target interaction (DTI) prediction, which inherently suffers from severe data imbalance, particularly in the class of true interactions, can benefit from oversampling techniques such as ROS and undersampling methods like RUS to enhance model accuracy and improve the detection of real interactions(Hasanin et al., 2019) . Therefore, the strategic use of these techniques in DTI prediction is not only reasonable but also aligned with their proven success in other medical applications.

Building on the success of resampling techniques in general healthcare applications, Chen et al. proposed a novel computational model for predicting drug–protein interactions, incorporating dimensionality reduction through Random Projection, data balancing via the NearMiss (NM) method, and model training using the Random Forest algorith(F. Chen et al., 2025)m.Khojasteh et al. introduced a novel multi-step approach called SRX-DTI for predicting drug–protein interactions, which employs diverse protein descriptors and FP2 drug fingerprints for feature extraction, utilizes the One-SVM-US method to address data imbalance, and applies the XGBoost algorithm for model training(Khojasteh et al., 2023).Liyaqat et al. proposed a method for predicting drug–protein interactions, utilizing PSSM for protein feature extraction, PubChem fingerprints for drug features, the NearestCUS technique to address class imbalance, and the CatBoost algorithm for model training(Liyaqat & Ahmad, 2023). Puri et al. proposed a new hybrid model for predicting drug–protein interactions (DTI), utilizing AM-PseAAC for protein representation and MSF for drug features. To address class imbalance, they applied the SMOTE-ENN resampling technique, and employed a combination of Random Forest and XGBoost algorithms for model training (Puri et al., 2022).

3. MATERIALS AND METHODS

In this study, an advanced method for predicting drug-target interactions (DTI) is introduced. First, the chemical structures of drugs in SMILE format and protein sequences in FASTA format are obtained from reliable sources using unique access identifiers. Then, different feature extraction techniques are used to identify unique characteristics from both drug compounds and protein sequences. These extracted features, along with known interaction data, are used to create drug-target pair vectors.

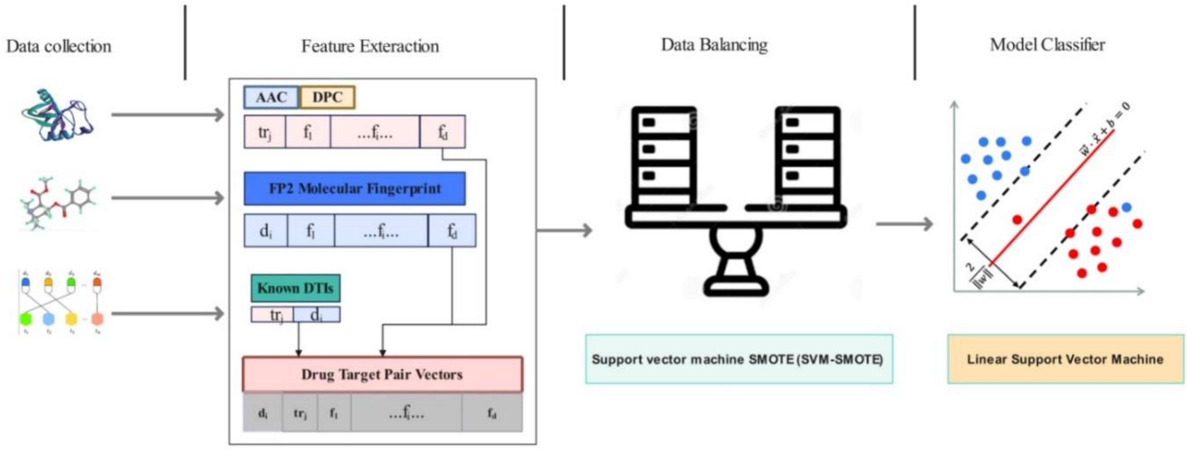
In the preprocessing stage, the data is first normalized using the MinMaxScaler method to ensure a consistent scale for all features. Then, principal component analysis (PCA) is applied to process high-dimensional feature vectors. This technique reduces the dimensionality of the data while preserving essential information, preventing irrelevant or redundant features from negatively impacting subsequent tasks.

To address the issue of class imbalance in the DTI dataset, we employed the **SVM-SMOTE** technique, which creates synthetic minority class samples close to the support vectors, thereby emphasizing the decision boundary. This approach enhances the model's generalization ability, particularly in DTI prediction, where positive samples are significantly underrepresented.

After balancing and refining the dataset, a Linear Support Vector Machine (LSVM) classifier is trained on the final dataset. A comprehensive evaluation was conducted to identify the most suitable learning algorithm, considering methods such as Nearest Neighbors, Linear SVM, Decision Tree, Random Forest, Neural Network, AdaBoost, Naive Bayes, QDA, LDA, EEC, RSC, and BBC. Among these algorithms, the Linear SVM classifier demonstrated the best and most reliable performance and was selected for training the model.

All stages, including data preprocessing, feature extraction, data balancing, and classification, have been implemented in an integrated system. This system optimizes the performance of the proposed model and provides a robust approach for predicting drug-target interactions. Figure 1 shows a diagram depicting the structure of our proposed model.

Figure 1. The Process Flow of The Proposed Model for Predicting Drug-Target Interactions.



3.1. Drug–Target Datasets

In this study, the drug-target interaction datasets were obtained from the collections curated by Yamanishi et al. (Yamanishi et al., 2008) . These datasets were sourced from reputable and publicly accessible databases such as KEGG BRITE (Kanehisa et al., 2012; Schomburg et al., 2004) , BRENDA(Schomburg et al., 2004) , SuperTarget (Günther et al., 2008) , and DrugBank (Wishart et al., 2006) , and are regarded as the "gold standard" in the field. The gold standard dataset is widely recognized for its reliability and is considered the most authoritative reference for evaluating other datasets. It includes four distinct categories of human drug-target interaction networks: enzymes (EN), ion channels (IC), G-protein-coupled receptors (GPCRs), and nuclear receptors (NR). The known interactions within these categories are 2,926 for enzymes, 1,476 for ion channels, 635 for GPCRs, and 90 for nuclear receptors. There is a notable relationship observed between the structural similarity of drugs, the sequence similarity of target proteins, and the interaction network topology. A summary of the gold standard dataset used in this research is presented in Table 1. The datasets can be publicly accessed at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

Table 1: Summary Of Gold Standard Dataset

	EN	GPCR	IC	NR
Drug	445	223	210	54
Target	664	95	204	26
Interaction	2926	635	1476	90

3.2. Feature Extraction Methods

To identify drug-target interactions using machine learning algorithms, extracting statistical features from drugs and proteins can play a crucial role in improving the accuracy and efficiency of the models. These features, derived from the distribution of molecular structural and chemical characteristics, are used as input data for machine learning algorithms. Statistical features enable machine learning models to simulate the complex relationships and patterns between the characteristics of drugs and proteins, allowing for more accurate predictions of potential interactions. In other words, these features assist algorithms in extracting meaningful patterns and useful information from raw data, thereby enhancing the accuracy of drug-protein interaction prediction models in simulation and drug screening processes.

In the final step of feature extraction, a total of 256 statistical features were extracted from the drugs, 20 features from AAC, and 400 features from the DPC of target proteins. As a result, after completing this process, a combined dataset with 676 features was obtained, encompassing both drug and target-related features. All extracted features are organized into a matrix, where the columns represent the features of both the drug and target protein, and the rows correspond to different drug-target protein samples. The last column of the matrix is dedicated to indicating known and unknown interactions. Details regarding the number of known and unknown interactions and datasets size in each dataset are presented in Table 2.

Table 2: Summary of Known Interaction - Unknown Interaction and Dataset Size in each Datasets

	Known Interaction	Unknown interaction	Datasets Size
NR	90	1314	**1404*678**
IC	1476	41364	**42840*678**
GPCR	635	20550	**21185*678**
EN	2926	292554	**295480*678**

3.2.1. Drug features

With the continuous progress in medicinal and synthetic organic chemistry, the variety of drug molecules has grown significantly. Cheminformatics, through the use of molecular fingerprints, enables quick comparisons and plays a pivotal role in structure-activity relationship (SAR) studies and virtual screening processes(Caron et al., 2020; Naveja & Medina-Franco, 2017; Naveja & Medina-Franco, 2017). Molecular fingerprints are binary representations that capture a range of characteristics with different complexities, such as the count of hydrogen atoms or the presence of specific molecular substructures(Alpay et al., 2022). Within the pharmaceutical domain, descriptors like FP2, FP3, FP4, and MACCS are commonly employed to represent the molecular structure of drugs(Dong et al., 2018). In this research, the FP2 format is chosen to depict the pharmaceutical compounds. The process of extracting a molecular fingerprint for a drug involves several steps: Initially, a Mol file containing the chemical structure of the drug is obtained using the drug's code from the KEGG database (Source: http://www.kegg.jp/kegg/). Then, the open-source software OpenBabel (downloadable from [http://openbabel.org/]) is used to convert and process the chemical data in various formats. Finally, the Mol files of the drug are transformed into the FP2 molecular fingerprint format using OpenBabel software. The resulting FP2 fingerprint is a 256-character hexadecimal string, which can be converted into decimal values ranging from 0 to 15, resulting in a 256-dimensional vector that represents the drug molecule (Shi et al., 2019a).

3.2.2. Target features

A) AAC (Amino acid composition (AAC))

Amino acid composition (AAC) is an important characteristic in the analysis of protein sequences, focusing on the frequency of each of the 20 standard amino acids found in proteins. This feature is represented as a 20-dimensional vector, in which each dimension shows the occurrence rate of a specific amino acid. The 20 amino acids included in this composition are ACDEFGHIKLMNPQRSTVWY(Charoenkwan et al., 2021b; Guo et al., 2020; Sun et al., 2020; D. Wang et al., 2011). The frequency of the occurrence of each amino acid is calculated using the following formula:

$$f_t = \frac{N(t)}{N} \ \ , \ t \in \{A, C, D, \dots, Y\} \qquad (1)$$

In this formula (1), N(t) explains the number of occurrences of amino acid type t in the protein sequence, while N stands for the total length of the protein sequence. AAC has been proven to be effective for identifying distinct patterns within proteins and is an essential feature in predictive tasks, such as differentiating thermophilic proteins from mesophilic ones. The use of tools like iLearnPlus for feature extraction and model building based on such compositions has been demonstrated in various research works(Ai et al., 2018; Guo et al., 2020) .

B) Dipeptide composition (DPC):

Dipeptide Composition (DPC) is a technique employed to calculate the frequency of dipeptide pairs within peptide sequences without considering the specific order of the amino acids. This method is highly valuable in protein sequence analysis, particularly for homologous data, and can generate as many as 400 distinct features to balance dipeptide occurrences. The calculation of DPC follows this formula:

$$DPC(s,t) = \frac{N(s,t)}{L-1} \ \ , \ s,t \ \in \{A, C, D, \dots, Y\} \qquad (2)$$

In this formula (2), N(s,t) refers to the count of dipeptides formed from the amino acid pair s and t, and L defines the length of the peptide sequence(Z. Chen et al., 2018; Saravanan & Gautham, 2015). Utilizing this feature improves the accuracy and efficiency of the analysis, transforming peptide sequences into meaningful numerical representations. Additionally, amino acids are divided into different groups based on their physical and chemical properties, such as aliphatic (G1: A, G, L, I, M, V), aromatic (G2: F, W, Y), positively charged (G3: H, R, K), negatively charged (G4: D, E), and uncharged (G5: C, N, P, Q, S, T) groups, which are incorporated as supplementary factors in the analysis(T. Y. Lee et al., 2011). This methodology is widely used in fields such as drug discovery, protein engineering, and identifying protein functional regions, ultimately providing valuable insights into protein functionality and biological activities.

3.3. Data Balancing Technique

Data imbalance biases machine learning models toward predicting non-interacting pairs, reducing their accuracy in identifying true interactions. Therefore, developing data balancing methods and more accurate models is crucial for improving drug-target interaction prediction.(Bekkar et al., 2013; Ezzat et al., 2016) . To address the class imbalance problem, three approaches have been developed: the cost-sensitive approach, the algorithm-based approach, and the data-based approach. In the cost-sensitive approach, misclassification costs are adjusted based on the importance of the class and the degree of data imbalance. In the algorithm-based approach, classification algorithms are modified to account for the issue of data imbalance (Elreedy & Atiya, 2019). The data-based approach aims to balance the classes by modifying the data distribution, which is achieved through under-sampling or over-sampling techniques. Under-sampling involves removing less important samples from the majority class, which potentially leads to the loss of valuable information. On the other hand, over-sampling, which is typically more effective, generates new data for the minority class. One of the successful over-sampling methods is SMOTE (Synthetic Minority Over-sampling Technique) (Yin et al., 2022). Consequently, integrating SMOTE with other model adjustment strategies, such as feature selection, class-weight adjustment, and deep learning algorithms, can enhance model performance and improve the accuracy of drug-target interaction predictions(Li et al., 2024)

In this study, the SVM-SMOTE (Support Vector Machine-Synthetic Minority Over-sampling Technique) method has been used to balance the dataset. SVM-SMOTE is an advanced version of SMOTE that utilizes Support Vector Machines (SVM) to enhance the sampling process and generate synthetic data. Unlike standard SMOTE, which creates new samples without considering decision boundaries, SVM-SMOTE leverages support vectors to generate synthetic samples in regions that have the greatest impact on class separation. In this method, the SVM algorithm is first trained on the training data to determine the decision boundary between classes (Fiifi et al., 2024). Then, new samples from the minority class are generated near this boundary. This process not only improves the quality of synthetic data but also reduces data overlap and mitigates the risk of overfitting by focusing on critical decision-making regions (Azlim Khan & Ahamed Hassain Malim, 2023b). Placing new samples near decision boundaries makes these boundaries more precise and well-defined, allowing the model to better distinguish between classes. This approach improves the model's performance, especially when dealing with imbalanced data and cases where class boundaries are complex and non-linear(H. Aljawazneh, 2021).

Imagine that SV_i is a support vector from the minority class. To generate a new synthetic sample, the nearest k neighbors of SV_i from the same minority class are first identified, and one of these neighbors is selected as SV_{ni}. Then, a new synthetic sample, SV_{new}, is created using the following formula:

$$SV_{new} = SV_i + \lambda \times (SV_{ni} + SV_i) \tag{3}$$

In this formula (3), λ is a random number between 0 and 1. This formula ensures that the new synthetic sample, SV_{new}, lies along the linear segment between SV_i and SV_{ni}, which helps strengthen the decision boundary between classes(Zheng, 2020).

In summary, SVM-SMOTE emphasizes generating synthetic data near support vectors, which ensures that the model's decision boundaries become more accurate and well-defined. This targeted approach helps the model perform better, especially when the data is imbalanced or when class boundaries are not easily distinguishable.

3.4. Classification Method

In machine learning, binary pattern separation is one of the most important tasks, which involves classifying observations into two distinct classes. This challenge is applicable in many fields such as robotics, environmental engineering, medical image analysis, and computer security. Methods such as decision trees, logistic regression, and nearest neighbors are used for this task, but Support Vector Machines (SVM) are regarded as one of the best methods due to their high accuracy and ability to separate complex data (Faccini et al., 2022). The goal of the SVM algorithm is to identify a hyperplane (decision boundary) that effectively separates the data into two distinct classes. In a two-dimensional space, the hyperplane is a line that divides the data points. The SVM algorithm seeks to choose this line in such a way that the distance between the data points and the line is maximized. This distance is referred to as the margin, and the objective of SVM is to enlarge this margin as much as possible to improve classification accuracy. During the learning process, SVM identifies the support vectors—data points that are closest to the decision boundary and have a significant influence on the positioning of the hyperplane. These points act as boundary markers, and the model strives to find an optimal hyperplane using them(Herle et al., 2020). One of the most commonly used versions of this algorithm is the Linear Support Vector Machine (Linear SVM), which divides data into two separate classes using a line or hyperplane, attempting to maximize the margin between these classes(Jailani et al., 2022).

In drug-target interaction (DTI) prediction, the Support Vector Machine (SVM) is a strong tool that identifies potential interactions by constructing hyperplanes or decision boundaries. This algorithm creates a hyperplane that divides the data into two sections, effectively separating interacting and non-interacting pairs. Essentially, SVM aims to find a boundary that increases the separation between data points of the two classes (interaction and non-interaction). With this capability, SVM can effectively recognize complex interactions present in biological data (Bagherian, Sabeti, et al., 2021; Lo et al., 2018).One of the key reasons for the widespread use of SVM in DTI prediction is its high accuracy in identifying real interactions and its ability to handle complex and noisy data. In addition to its high precision, SVM has a relatively low computational cost, making it highly suitable for large-scale and complex analyses. As a result, SVM is widely employed as an effective method for drug-target interaction prediction in various biomedical and pharmaceutical studies (Xu et al., 2021).

4. RESULT

This section addresses the evaluation and optimization of the proposed model for predicting drug–target interactions (DTIs). All modeling steps—including feature extraction, data balancing, and model training—were implemented using Python (version 3.10) and established libraries such as Scikit-learn and TensorFlow. To address class imbalance, the SVM-SMOTE method was employed with default settings, including sampling_strategy='auto', k_neighbors=5, m_neighbors=10, out_step=0.5, and a default RBF SVM as the internal classifier. To assess model performance, we employed 5-fold cross-validation, ensuring that all data were equally utilized for both training and testing by randomly partitioning the dataset in each iteration. Additionally, to optimize the SVM classifier, we conducted a grid search over a defined parameter space (e.g., $C \in [0.1, 100]$ and $\gamma \in [0.001, 1]$). Model performance was evaluated based on the AUC score, and the parameter configuration yielding the highest average AUC across all folds was selected.We evaluated the proposed model on four benchmark datasets—Enzyme, Ion Channel, Nuclear Receptor, and GPCR—and compared its performance against a variety of machine learning algorithms, sampling strategies, and other leading DTI prediction approaches.

4.1. Evaluation of SVM-SMOTE-Based Models Using ROC and Precision-Recall Curves

Figures 2 and 3 present the ROC and Precision-Recall curves used to evaluate the performance of various machine learning models in predicting drug-target interactions across four benchmark datasets: EN, GPCR, IC, and NR. These curves are powerful tools for assessing a model's ability to distinguish between positive and negative samples, and their interpretation provides valuable insights into both the accuracy and discriminative capacity of the models. The Linear SVM demonstrated consistently strong and stable performance among the evaluated models. In most cases, it achieved high Precision and Recall while maintaining a favorable balance between them. This robustness is largely attributed to the SVM's capacity to identify optimal decision boundaries, particularly in high-dimensional feature spaces and complex biological data.

In the ROC plots, both the Random Forest and Linear SVM models exhibited very high Area Under the Curve (AUC) values across all datasets, indicating near-ideal performance. An AUC close to 1 reflects a model's strong ability to accurately distinguish between positive and negative classes. The selection of Linear SVM as the final model in this study was based on its stability and generalizability. While Random Forest achieved perfect AUC and F1 scores (1.00) in several datasets, these results may indicate potential overfitting. Due to its ensemble nature, Random Forest can become overly tailored to the training data, especially when the data contains repetitive patterns or dominant features. In such cases, its performance may deteriorate on unseen data. In contrast, the SVM model, with its linear and margin-maximizing approach, tends to maintain a better balance between accuracy and generalizability, making it more reliable for real-world applications.

In contrast, models such as Naïve Bayes and AdaBoost typically exhibited lower AUC values, especially on more complex datasets like IC and NR. The decreased performance of Naïve Bayes is often attributed to its assumption of feature independence, which rarely holds true in biological data. Additionally, AdaBoost frequently struggled to generalize well due to its high sensitivity to noise and imbalanced data.

A similar trend is evident in the Precision-Recall plots. Models such as Random Forest, Linear SVM, and to some extent Neural Networks and BBC have demonstrated a commendable balance between Precision and Recall. Achieving this balance is crucial for accurately identifying true positives while minimizing the occurrence of false positives. In this regard, Linear SVM—whose performance approaches the ideal—stands out as a dependable choice for modeling under complex and imbalanced conditions.

Conversely, models like RSC, QDA, and EEC performed less favorably, likely due to their limited capacity to handle nonlinear relationships and heterogeneous feature sets typical of biological data. This limitation is particularly pronounced in smaller datasets, such as NR. Specifically, QDA's assumption of normal distribution within each class often undermines its effectiveness when dealing with such data.

In the IC dataset, which contains information about ion channels, models such as AdaBoost and RSC exhibited a significant decline in performance. This highlights the difficulty these models face in capturing the complex and nonlinear patterns inherent in biological data. In contrast, models like Linear SVM and Random Forest demonstrated more stable performance by leveraging greater flexibility or structural robustness. Overall, the application of the SVM-SMOTE method for data balancing played a crucial role in enhancing the accuracy of most models and, notably, provided a strong foundation for the Linear SVM model to achieve precise drug-target interaction detection while maintaining its generalizability.

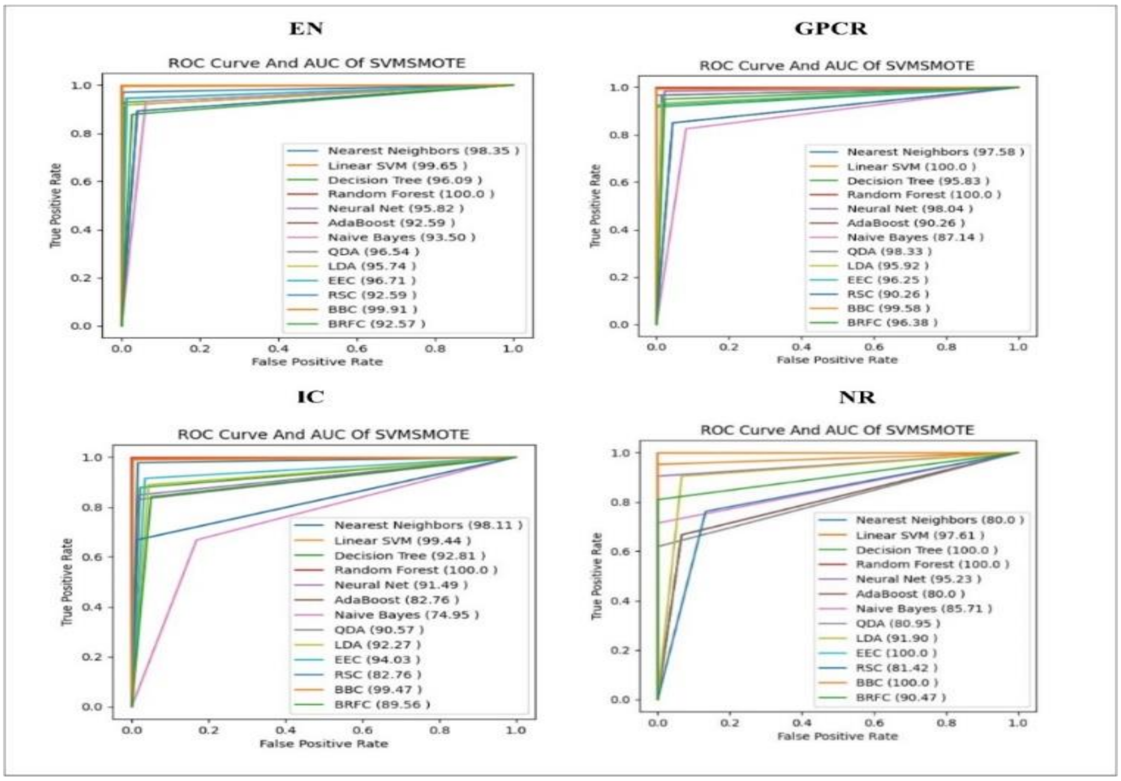


Figure 2. ROC Curves of SVM-SMOTE Models on EN, GPCR, IC, and NR Datasets

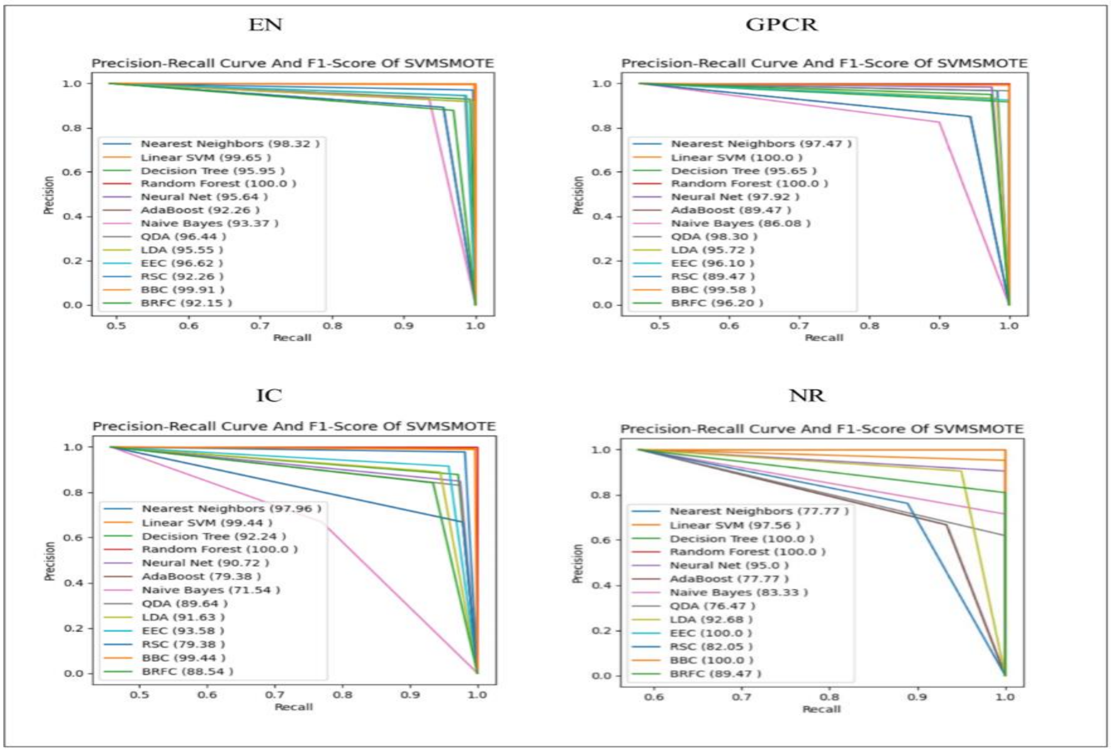


Figure 3. Precision-Recall Curves of SVM-SMOTE Models on EN, GPCR, IC, and NR Datasets

4.2. Comparison of Machine Learning Models Enhanced with SVM-SMOTE on Four Different Data Sets

In this section, the performance of machine learning models improved with the SVM-SMOTE technique for predicting drug-target interactions is evaluated on four different datasets. The analysis shows that some models performed well on all datasets, while other models struggled to identify positive interactions. The performance of different SVM-SMOTE-based machine learning methods in DTI prediction is shown in Table 3.

In the analysis of the Enzyme (EN) dataset, the Random Forest and Linear SVM models performed best in predicting enzyme-protein interactions. Random Forest achieved 100% accuracy across all metrics, while Linear SVM performed well with an F1 score of 0.996. The BBC model also showed high accuracy. The neural network model had a lower recall (0.916), indicating that it missed some positive interactions. The AdaBoost, RSC, and BRFC models performed worse, with their lower recall indicating challenges in identifying positive examples.

In the analysis of the G-protein-coupled receptors (GPCRs) dataset, the random forest and linear SVM models performed best, achieving 100% accuracy. The BBC model also showed strong performance with an F1 score of 0.995. The QDA model performed well, with an F1 score of 0.983. In contrast, the Naive Bayes, AdaBoost, and RSC models performed poorly. In particular, Naive Bayes had a recall value of 0.825, indicating its weakness in identifying real interactions. The neural network model also performed well but was slightly weaker compared to the top models.

In the analysis of the ion channel (IC) dataset, the Random Forest, Linear SVM, and BBC models performed very well, showing high generalization ability and achieving F1 scores close to 1.00. In contrast, the Naive Bayes, AdaBoost, RSC, and QDA models performed poorly in identifying positive interactions. The LDA model showed stable performance with an F1 score of 0.916. The neural network model also achieved a good balance, but its low recall value (0.848) indicates difficulty in identifying some real interactions.

In the analysis of the nuclear receptor (NR) dataset, the random forest, decision tree, EEC, and BBC models achieved 100% accuracy across all metrics, indicating strong performance, although overfitting may be a concern. In contrast, the KNN, QDA, and AdaBoost models had low recall and did not effectively identify true interactions. The linear SVM achieved a good balance between precision and generalization with an F1 score of 0.975. The neural network model also performed well, but its low recall value (0.904) suggests that some interactions were missed.

Table 3. Performance of Machine Learning Models Enhanced with SVM-SMOTE in Drug-Target Interaction Prediction

Dataset	*Machine Learning Models*	*model's performance*				
		precision	**recall**	**f1-score**	**accuracy**	**ROC**
EN	Nearest Neighbors	0.996429	0.970435	0.98326	0.983775	0.98354
	Linear SVM	0.998255	0.994783	0.996516	0.996584	0.996552
	Decision Tree	0.992565	0.928696	0.959569	0.961571	0.960992
	Random Forest	1	1	1	1	1
	Neural Net	1	0.916522	0.956443	0.959009	0.958261
	AdaBoost	0.955307	0.892174	0.922662	0.926558	0.925953
	Naive Bayes	0.935428	0.932174	0.933798	0.935098	0.935047
	QDA	0.985481	0.944348	0.964476	0.965841	0.965463
	LDA	0.998106	0.916522	0.955576	0.958155	0.957422
	EEC	0.987296	0.946087	0.966252	0.967549	0.967171
	RSC	0.955307	0.892174	0.922662	0.926558	0.925953
	BBC	1	0.998261	0.99913	0.999146	0.99913
	BRFC	0.96929	0.878261	0.921533	0.926558	0.925708
GPCR	Nearest Neighbors	0.983051	0.966667	0.97479	0.976378	0.975871
	Linear SVM	1	1	1	1	1
	Decision Tree	1	0.916667	0.956522	0.96063	0.958333
	Random Forest	1	1	1	1	1
	Neural Net	0.975207	0.983333	0.979253	0.980315	0.980473
	AdaBoost	0.944444	0.85	0.894737	0.905512	0.902612
	Naive Bayes	0.9	0.825	0.86087	0.874016	0.871455
	QDA	1	0.966667	0.983051	0.984252	0.983333
	LDA	0.982456	0.933333	0.957265	0.96063	0.959204
	EEC	1	0.925	0.961039	0.964567	0.9625
	RSC	0.944444	0.85	0.894737	0.905512	0.902612
	BBC	1	0.991667	0.995816	0.996063	0.995833
	BRFC	0.974359	0.95	0.962025	0.964567	0.963806
IC	Nearest Neighbors	0.981481481	0.977859779	0.979667283	0.981387479	0.981117389
	Linear SVM	1	0.988929889	0.994434137	0.994923858	0.994464945
	Decision Tree	0.971428571	0.878228782	0.92248062	0.932318105	0.928176891
	Random Forest	1	1	1	1	1
	Neural Net	0.974576271	0.848708487	0.90729783	0.920473773	0.914979244
	AdaBoost	0.978378378	0.667896679	0.793859649	0.840947547	0.827698339
	Naive Bayes	0.770212766	0.667896679	0.71541502	0.756345178	0.749573339
	QDA	0.974025974	0.830258303	0.896414343	0.912013536	0.905754151
	LDA	0.945098039	0.889298893	0.91634981	0.925549915	0.922774446
	EEC	0.957528958	0.915129151	0.935849057	0.942470389	0.940377076
	RS	0.978378378	0.667896679	0.793859649	0.840947547	0.827698339
	BBC	0.996296296	0.992619926	0.994454713	0.994923858	0.994747463
	BRFC	0.93442623	0.841328413	0.885436893	0.900169205	0.895664207
NR	Nearest Neighbors	0.933333333	0.666666667	0.777777778	0.777777778	0.8
	Linear SVM	1	0.952380952	0.975609756	0.972222222	0.976190476
	Decision Tree	1	1	1	1	1
	Random Forest	1	1	1	1	1
	Neural Net	1	0.904761905	0.95	0.944444444	0.952380952
	AdaBoost	0.933333333	0.666666667	0.777777778	0.777777778	0.8
	Naive Bayes	1	0.714285714	0.833333333	0.833333333	0.857142857
	QDA	1	0.619047619	0.764705882	0.777777778	0.80952381

LDA	0.95	0.904761905	0.926829268	0.916666667	0.919047619
EEC	1	1	1	1	1
RS	0.888888889	0.761904762	0.820512821	0.805555556	0.814285714
BBC	1	1	1	1	1
BRFC	1	0.80952381	0.894736842	0.888888889	0.904761905

4.3. Evaluating the generalizability and stability of the model in different data sets

To critically assess the generalizability and stability of the proposed model's performance across different datasets, experiments were conducted using five repetitions of 5-fold cross-validation (5-Fold CV) on four datasets: EN, GPCR, IC, and NR. Alongside the proposed model, several other machines learning algorithms, including Nearest Neighbors, Linear SVM, Decision Tree, Random Forest, Neural Network, AdaBoost, and Naive Bayes, were also evaluated under identical conditions. The results, based on the mean AUROC and standard deviation (Std), revealed no significant variance in the performance of the proposed model across the repeated runs, nor in comparison to the other models. These findings demonstrate the robustness and stability of the proposed model, indicating its resistance to data noise and fluctuations resulting from random data partitioning. Table 4 presents the performance of various classifiers with SVM-SMOTE evaluated using 5-fold cross-validation on the enzyme, ion channel, GPCR, and nuclear receptor datasets

Furthermore, a more detailed examination of model performance across each dataset revealed that, while the proposed method consistently achieved superior AUROC scores in all datasets, the other evaluated algorithms also generally yielded acceptable results. This observation suggests that the datasets employed are of high quality, enabling most algorithms to extract meaningful patterns effectively. Nevertheless, the consistently high stability of the proposed model across all datasets and cross-validation folds highlights not only its strong predictive capability but also its superior generalizability and robustness compared to the other models. These findings further emphasize the significance of the proposed method as an effective and reliable framework for predicting drug-target interactions.

Table 4. Performance (5-fold CV) of Different Classifiers with SVM-SMOTE on Enzyme, Ion Channel, GPCR and Nuclear Receptor Datasets.

Data Set	**5-Fold Cross-Validation**							
	ML	1	2	3	4	5	Mean	Std
EN	Linear SVM	0.9966	0.9965	0.9962	0.9968	0.9964	0.9965	0.0015
	Decision Tree	0.9620	0.9580	0.9635	0.9640	0.9570	0.9609	0.0050
	Neural Net	0.9600	0.9560	0.9595	0.9600	0.9555	0.9582	0.0027
	AdaBoost	0.9245	0.9280	0.9260	0.9300	0.9220	0.9259	0.0042
	Naive Bayes	0.9380	0.9390	0.9320	0.9380	0.9280	0.9350	0.0084
GPCR	Linear SVM	0.9947	0.9942	0.9945	0.9940	0.9955	0.9944	0.0057
	Decision Tree	0.9565	0.9600	0.9590	0.9660	0.9500	0.9583	0.0270
	Neural Net	0.9820	0.9810	0.9820	0.9870	0.9700	0.9804	0.0335
	AdaBoost	0.9015	0.9025	0.9050	0.9040	0.9000	0.9026	0.0033
	Naive Bayes	0.8705	0.8750	0.8700	0.8760	0.8640	0.8714	0.0075
IC	Linear SVM	0.9780	0.9760	0.9745	0.9770	0.9750	0.9761	0.0355
	Decision Tree	0.9365	0.9280	0.9240	0.9300	0.9220	0.9281	0.0060
	Neural Net	0.9180	0.9125	0.9160	0.9180	0.9100	0.9149	0.0112
	AdaBoost	0.8250	0.8300	0.8280	0.8340	0.8220	0.8276	0.0067
	Naive Bayes	0.7520	0.7520	0.7450	0.7580	0.7400	0.7495	0.0104
NR	Linear SVM	0.9760	0.9785	0.9750	0.9770	0.9940	0.9761	0.0355
	Decision Tree	0.9950	1	1	1	0.9900	0.9970	0.0044
	Neural Net	0.9565	0.9500	0.9550	0.9600	0.9400	0.9523	0.0380
	AdaBoost	0.8000	0.7950	0.8050	0.8100	0.7900	0.8000	0.0300
	Naive Bayes	0.8570	0.8540	0.8600	0.8650	0.8500	0.8571	0.0488

4.4. Comparative Analysis of SVM-SMOTE and Data Balancing Methods for DTI Prediction

In this study, a comprehensive comparative analysis was conducted between the proposed SVM-SMOTE method in drug-target interaction (DTI) prediction and other data balancing methods. The comparison covered a wide range of UnderSampling, OverSampling, Hybrid techniques as well as Adaptive Synthetic Sampling (ADASYN) method. The evaluations were performed on four reference DTI-related datasets and at each stage, the data were trained using the Linear SVM (LSVM) algorithm after applying different balancing methods. The performance of the models was measured based on F1 score and area under the ROC curve (ROC-AUC). The main objective of this analysis was to highlight the capabilities and potential advantages of the SVM-SMOTE method compared to conventional methods, especially in dealing with the challenges associated with data imbalance in DTI prediction tasks. Table 5 presents the performance evaluation of various sampling methods applied to four DTI datasets (NR, IC, GPCR, EN), assessed using ROC-AUC and F1-score metrics.

The results showed that the combined SVM-SMOTE method outperformed other techniques and achieved significant results across all datasets. This superiority stems from its use of the Support Vector Machine (SVM) algorithm to accurately identify the decision boundary between classes and generate synthetic samples in critical regions, where the likelihood of misclassification is highest. Unlike methods such as RandomOverSampler or SMOTE, which generate samples without considering class boundaries, SVM-SMOTE enhances model accuracy by producing more targeted and diverse data. Furthermore, in contrast to data deletion methods like ClusterCentroids or OneSideSelection—which risk discarding valuable information—SVM-SMOTE maintains the original data while strategically augmenting the minority class, leading to better overall performance.RandUnderSampler, as an undersampling method, has performed very closely to— and in some cases even on par with—advanced oversampling methods such as SMOTENC and SMOTEN. In contrast, neighborhood-based methods like EditNearNeighb and ReEditNearNeighb showed very poor results on the NR dataset (ROC = 0.63 and 0.64, respectively). Additionally, although ADASYN performed well on some datasets, it exhibited a significant drop in F1-score on the IC and GPCR datasets, performing worse than other oversampling techniques.

Table 5. Performance of Various Sampling Methods across Four DTI Datasets (NR, IC, GPCR, EN) Using ROC-AUC and F1-Score Metrics.

Sampling Method		EN		GPCR		IC		NR	
		F1-Score	ROC	F1-Score	ROC	F1-Score	ROC	F1-Score	ROC
UnderSampling Method	ClusterCentroids	0.9538	0.9536	0.9552	0.955	0.955	0.9551	0.945	0.94
	ConNearNeigh	0.9894	0.9892	0.978	0.9774	0.971	0.9701	0.922	0.915
	EditNearNeighb	0.9789	0.9785	0.961	0.96	0.969	0.9679	0.72	0.63
	ReEditNearNeighb	0.9788	0.9784	0.951	0.956	0.979	0.9789	0.73	0.64
	NeighbCleanRule	0.9859	0.9857	0.982	0.982	0.983	0.9836	0.8	0.755
	OneSideSelection	0.9755	0.9755	0.9752	0.975	0.971	0.9715	0.828	0.795
	RandUnderSampler	0.9956	0.9955	0.999	0.9992	0.9939	0.9939	0.974	0.973
OverSampling Method	BorderlineSMOTE	0.9865	0.9866	0.989	0.9892	0.9839	0.9839	0.964	0.963
	RandOverSampler	0.9756	0.9755	0.979	0.9792	0.9739	0.9739	0.974	0.973
	SMOTE	0.9656	0.9655	0.969	0.9692	0.9639	0.9639	0.964	0.963
	SMOTENC	0.9956	0.9955	0.999	0.9992	0.9939	0.9939	0.974	0.973
	SMOTEN	0.9956	0.9955	0.999	0.9992	0.9939	0.9939	0.974	0.973
	SVMSMOTE	0.9965	0.9966	0.9993	0.9995	0.995	0.9949	0.977	0.977
Hybrid	SMOTETomek	0.9956	0.9955	0.992	0.9915	0.9939	0.9939	0.828	0.795
Adaptive Synthetic Sampling	ADASYN	0.9542	0.9536	0.970	0.976	0.9230	0.9962	0.88	0.9706

4.5. Comparison of SVM-SMOTE-Based Model with Other Methods

The aim of this section is to compare our proposed method with other existing approaches in the field of drug-target interaction (DTI) prediction. To this end, we evaluated four different methods using the AUC metric across four standard benchmark datasets, with the results presented in Table 6. All these studies primarily focus on data balancing and enhancing the accuracy of DTI prediction. Mahmud et al. introduced the iDTi-CSsmoteB model, which leverages protein sequence features and the chemical structure of drugs. To address data imbalance, they employed the SMOTE technique and used the XGBoost algorithm for classification(Mahmud et al., 2019). Chen et al. applied a variety of descriptors and utilized Random Projection for dimensionality reduction and NearMiss for data balancing, followed by classification using the Random Forest algorithm (F. Chen et al., 2025). Mahmud et al. proposed the pdti-EssB model, which integrates chemical structure features of drugs with sequence-based, structural, and evolutionary information of proteins. They applied undersampling techniques for data balancing and trained the final model using XGBoost (Mahmud et al., 2020). Lastly, Shi et al. developed the LRF-DTIs method, which demonstrated highly accurate DTI prediction performance by extracting features using PsePSSM and FP2, reducing dimensionality via the Lasso method, balancing data with SMOTE, and applying Random Forest for classification(Shi et al., 2019).

Table 6 compares the performance of five drug-target interaction (DTI) prediction methods, showing that the proposed method has a very impressive performance compared to other methods. This method recorded the best possible result in the GPCR dataset with an AUC of 1.0000, and in the NR dataset with an AUC of 0.9761, it had the highest performance among competitors. Also, in the EN and IC datasets, it had a performance very close to the best existing method (Shi et al., with values of 0.9965 and 0.9944), which indicates the high stability and generalizability of the model. This superiority is the result of using a precise combination of structural, sequence, and evolutionary features of proteins, optimal use of molecular fingerprints of drugs, the use of effective data balancing techniques, and finally the use of a powerful classification algorithm such as LSVM. Therefore, the proposed method not only provides better performance numerically, but also provides a comprehensive, accurate, and adaptable approach to common challenges in predicting drug-target interactions in terms of technical design.

Table 6. A Comparison of The Proposed Model with Existing Methods across Four Datasets.

drug-target interaction methods	**Dataset**			
	EN	GPCR	IC	NR
Mahmud et al. (2019)	0.9534	0.8797	0.9320	0.8350
F. Chen et al.(2025)	0.9933	0.9765	0.9821	0.9226
Mahmud et al. (2020)	0.9234	0.8797	0.9220	0.9226
Shi et al. (2019)	0.9982	0.9918	0.9965	0.9559
Proposed method	0.9965	1	0.9944	0.9761

5. CONCLUSION

In this study, an effective framework for drug–target interaction (DTI) prediction was proposed, achieving strong performance across four benchmark datasets—Enzyme, GPCR, Ion Channel, and Nuclear Receptor. The model integrates protein features (AAC and DPC), FP2 molecular fingerprints, and the SVM-SMOTE technique for addressing class imbalance. Trained using the Linear Support Vector Machine (LSVM) algorithm, the proposed approach demonstrated competitive, and in many cases superior, performance compared to commonly used methods.

One of the key distinguishing aspects of this framework is the integration of LSVM with SVM-SMOTE, a resampling technique that is aware of the decision boundary. Unlike conventional methods such as SMOTE or ADASYN, SVM-SMOTE leverages the geometric structure of the SVM model to generate more realistic minority class samples, thereby improving data balance without distorting the original feature space. Additionally, LSVM, as a model based on margin maximization and convex optimization, is theoretically guaranteed to converge to the optimal solution. The presence of the C parameter also allows for precise control over the trade-off between training error and model complexity. Together, these characteristics lead to greater robustness to noise, reduced risk of overfitting, enhanced interpretability, and improved computational efficiency compared to models like Random Forest.

However, it is worth noting that on certain datasets, the Random Forest algorithm outperformed in terms of metrics such as AUC and F1-score. This suggests that in scenarios involving complex patterns or pronounced nonlinear relationships, tree-based models may offer distinct advantages. Nevertheless, LSVM was selected as the primary method due to its theoretical strengths—such as margin maximization and convex optimization—as well as its simplicity, computational efficiency, interpretability, and robustness to noise.

Beyond its predictive performance, the proposed model holds practical value in the early stages of the drug discovery pipeline. It can serve as a pre-screening tool to efficiently filter potential drug candidates before moving to costly and time-consuming experimental validation. Given its high recall and AUC scores, the model is particularly well-suited to identifying true positive interactions, including rare or previously unknown drug-target pairs. Furthermore, the availability of open-source code and datasets allows researchers to fine-tune or retrain the model on specific targets or chemical spaces, making it a flexible and useful tool in real-world biomedical research and pharmaceutical development.

Overall, the proposed framework demonstrated strong performance in predicting drug–target interactions. However, certain limitations remain, including the absence of evaluation on real-world datasets, limited biological diversity, and a reliance on hand-crafted features. Future research could address these challenges by incorporating multi-omics data, applying transfer learning techniques, and leveraging graph-based models to enhance both accuracy and generalizability. predictions, such as identifying key amino acid or dipeptide features and analyzing their biological relevance in drug–target interactions. This would significantly enhance the model's utility and interpretability for biomedical researchers.

REFERENCES

Abbasi, K., Razzaghi, P., Poso, A., Amanlou, M., Ghasemi, J. B., & Masoudi-Nejad, A. (2020). DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. *Bioinformatics*, *36*(17), 4633-4642.

Ai, H., Zhang, L., Zhang, J., Cui, T., Chang, A. K., & Liu, H. (2018). Discrimination of thermophilic and mesophilic proteins using support vector machine and decision tree. *Current Proteomics*, *15*(5), 374-383.

Aljawazneh, H., Mora, A. M., García-Sánchez, P., & Castillo-Valdivieso, P. A. (2021). Comparing the performance of deep learning methods to predict companies' financial failure. *IEEE Access*, *9*, 97010-97038.

Alpay, B. A., Gosink, M., & Aguiar, D. (2022). Evaluating molecular fingerprint-based models of drug side effects against a statistical control. *Drug Discovery Today*, *27*(11), 103364.

An, Q., & Yu, L. (2021). A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Briefings in bioinformatics*, *22*(6), bbab275.

Atta Mills, E. F. E., Deng, Z., Zhong, Z., & Li, J. (2024). Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques. *Journal of Big Data*, *11*(1), 170.

Azlim Khan, A. K., & Ahamed Hassain Malim, N. H. (2023). Comparative studies on resampling techniques in machine learning and deep learning models for drug-target interaction prediction. *Molecules*, *28*(4), 1663.

Bagherian, M., Kim, R. B., Jiang, C., Sartor, M. A., Derksen, H., & Najarian, K. (2021). Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions. *Briefings in bioinformatics*, *22*(2), 2161-2171.

Bagherian, M., Sabeti, E., Wang, K., Sartor, M. A., Nikolovska-Coleska, Z., & Najarian, K. (2021). Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics*, *22*(1), 247-269.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, *3*(10).

Bian, J., Zhang, X., Zhang, X., Xu, D., & Wang, G. (2023). MCANet: shared-weight-based MultiheadCrossAttention network for drug–target interaction prediction. *Briefings in Bioinformatics*, *24*(2), bbad082.

Caron, G., Digiesi, V., Solaro, S., & Ermondi, G. (2020). Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discovery Today*, *25*(4), 621-627.

Charoenkwan, P., Chotpatiwetchkul, W., Lee, V. S., Nantasenamat, C., & Shoombuatong, W. (2021). A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Scientific Reports*, *11*(1), 23782.

Chen, F., Zhao, Z., Ren, Z., Lu, K., Yu, Y., & Wang, W. (2025). Prediction of drug target interaction based on under sampling strategy and random forest algorithm. *PloS one*, *20*(3), e0318420.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., ... & Song, J. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, *34*(14), 2499-2502.

Dong, J., Yao, Z. J., Zhang, L., Luo, F., Lin, Q., Lu, A. P., ... & Cao, D. S. (2018). PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *Journal of cheminformatics*, *10*, 1-11.

El-Behery, H., Attia, A. F., El-Fishawy, N., & Torkey, H. (2022). An ensemble-based drug–target interaction prediction approach using multiple feature information with data balancing. *Journal of Biological Engineering*, *16*(1), 21.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32-64.

Ezzat, A., Wu, M., Li, X., & Kwoh, C. K. (2018). Computational prediction of drug-target interactions via ensemble learning. In *Computational methods for drug repurposing* (pp. 239-254). New York, NY: Springer New York.

Ezzat, A., Wu, M., Li, X. L., & Kwoh, C. K. (2016). Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC bioinformatics*, *17*, 267-276.

Faccini, D., Maggioni, F., & Potra, F. A. (2022). Robust and distributionally robust optimization models for linear support vector machine. *Computers & Operations Research*, *147*, 105930.

Faulon, J. L., Misra, M., Martin, S., Sale, K., & Sapra, R. (2008). Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics*, *24*(2), 225-233.

Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., & Zhang, P. (2018, July). Interpretable drug target prediction using deep neural representation. In *IJCAI* (Vol. 2018, pp. 3371-3377).

Gao, S., Liu, Z., & Li, Y. (2022). Networks and algorithms in heterogeneous network-based methods for drug-target interaction prediction: A survey and comparison. In *Proceedings of the 1st International Conference on Health Big Data and Intelligent Healthcare*.

Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., ... & Preissner, R. (2007). SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic acids research*, *36*(suppl_1), D919-D922.

Guo, Z., Wang, P., Liu, Z., & Zhao, Y. (2020). Discrimination of thermophilic proteins and non-thermophilic proteins using feature dimension reduction. *Frontiers in Bioengineering and Biotechnology*, *8*, 584807.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced big data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1), 1-25.

Herle, A., Channegowda, J., & Prabhu, D. (2020, July). Quasar detection using linear support vector machine with learning from mistakes methodology. In *2020 IEEE international conference on electronics, computing and communication technologies (CONECCT)* (pp. 1-6). IEEE.

Hu, S., Xia, D., Su, B., Chen, P., Wang, B., & Li, J. (2019). A convolutional neural network system to discriminate drug-target interactions. *IEEE/ACM transactions on computational biology and bioinformatics*, *18*(4), 1315-1324.

Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2021). MolTrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, *37*(6), 830-836.

Huang, M.-W., Chiu, C.-H., Tsai, C.-F., & Lin, W.-C. (2021). On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer Prediction. *Applied Sciences*, *11*(14), 6574. https://doi.org/10.3390/app11146574

Ikechukwu, D., & Kumar, A. (2023). Drug-Target-Interaction Prediction with Contrastive and Siamese Transformers. *bioRxiv*, 2023-10.

Jailani, N. S. J., Muhammad, Z., Rahiman, M. H. F., & Taib, M. N. (2022). Intelligent grading of kaffir lime oil quality using non-linear support vector machine. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(6), 6716-6723.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, *40*(D1), D109-D114.

Khojasteh, H., Pirgazi, J., & Ghanbari Sorkhi, A. (2023). Improving prediction of drug-target interactions based on fusing multiple features with data balancing and feature selection techniques. *Plos one*, *18*(8), e0288173.

Latief, M. A., Nabila, L. R., Miftakhurrahman, W., Ma'rufatullah, S., & Tantyoko, H. (2024). Handling Imbalance Data using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification. *Int. J. Eng. Comput. Sci. Appl*, *3*(1), 11-18.

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, *15*(6), e1007129.

Lee, T. Y., Chen, S. A., Hung, H. Y., & Ou, Y. Y. (2011). Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PloS one*, *6*(3), e17331.

Li, Y., Cui, X., Yang, X., Liu, G., & Zhang, J. (2024). Artificial intelligence in predicting pathogenic microorganisms' antimicrobial resistance: challenges, progress, and prospects. *Frontiers in Cellular and Infection Microbiology*, *14*, 1482186.

Liyaqat, T., & Ahmad, T. (2023). A machine learning strategy with clustering under sampling of majority instances for predicting drug target interactions. *Molecular Informatics*, *42*(5), 2200102.

Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, *23*(8), 1538-1546.

Madhukar, N. S., Khade, P. K., Huang, L., Gayvert, K., Galletti, G., Stogniew, M., ... & Elemento, O. (2019). A Bayesian machine learning approach for drug target identification using diverse data types. *Nature communications*, *10*(1), 5221.

Mahmud, S. H., Chen, W., Jahan, H., Liu, Y., Sujan, N. I., & Ahmed, S. (2019). iDTi-CSsmoteB: identification of drug–target interaction based on drug chemical structure and protein sequence using XGBoost with over-sampling technique SMOTE. *IEEE Access*, *7*, 48699-48714.

Mahmud, S. H., Chen, W., Liu, Y., Awal, M. A., Ahmed, K., Rahman, M. H., & Moni, M. A. (2021). PreDTIs: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Briefings in bioinformatics*, *22*(5), bbab046.

Mitchell, J. B. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(5), 468-481.

Moesgaard, L. K. (2024). Understanding P-glycoprotein inhibition from a molecular basis–development of rational design strategies for P-glycoprotein inhibitors.

Naveja, J. J., & Medina-Franco, J. L. (2017). ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Research*, *6*, Chem-Inf.

Padhi, A., Agarwal, A., Saxena, S. K., & Katoch, C. D. S. (2023). Transforming clinical virology with AI, machine learning and deep learning: a comprehensive review and outlook. *VirusDisease*, *34*(3), 345-355.

Prasetyo, V. P., & Anggraeni, W. (2024, August). Drug-Target Interactions Prediction Using Stacking Ensemble Learning Approach. In *2024 International Electronics Symposium (IES)* (pp. 681-686). IEEE.

Redkar, S., Mondal, S., Joseph, A., & Hareesha, K. S. (2020). A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing. *Molecular informatics*, *39*(5), 1900062.

Saravanan, V., & Gautham, N. (2015). Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. *Omics: a journal of integrative biology*, *19*(10), 648-658.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic acids research*, *32*(suppl_1), D431-D433.

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, *111*(6), 1839-1852.

Sorkhi, A. G., Abbasi, Z., Mobarakeh, M. I., & Pirgazi, J. (2021). Drug–target interaction prediction using unifying of graph regularized nuclear norm with bilinear factorization. *BMC bioinformatics*, *22*, 1-23.

Sun, S., Ao, C., Wang, D., & Dong, B. (2020). The frequencies of oppositely charged, uncharged polar, and β-branched amino acids determine proteins' thermostability. *IEEE Access*, *8*, 66839-66845.

Vlasiou, M. C. (2024). *Computer-Aided Drug Discovery Methods: A Brief Introduction*. Bentham Science Publishers.

Wang, D., Yang, L., Fu, Z., & Xia, J. (2011). Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction. *Protein and peptide letters*, *18*(7), 684-689.

Wang, L., Zhou, Y., & Chen, Q. (2023). Ammvf-dti: A novel model predicting drug–target interactions based on attention mechanism and multi-view fusion. *International Journal of Molecular Sciences*, *24*(18), 14142.

Wang, X. R., Cao, T. T., Jia, C. M., Tian, X. M., & Wang, Y. (2021). Quantitative prediction model for affinity of drug–target interactions based on molecular vibrations and overall system of ligand-receptor. *BMC bioinformatics*, *22*, 1-18.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, *34*(suppl_1), D668-D672.

Xu, L., Ru, X., & Song, R. (2021). Application of machine learning for drug–target interaction prediction. *Frontiers in genetics*, *12*, 680117.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., & Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13), i232-i240.

Yin, L., Du, X., Ma, C., & Gu, H. (2022). Virtual screening of drug proteins based on the prediction classification model of imbalanced data mining. *Processes*, *10*(7), 1420.

Zheng, X. (2020). *SMOTE variants for imbalanced binary classification: heart disease prediction*. University of California, Los Angeles.