CHAPTER 1

INTRODUCTION

1.1 Rationale

Drug Target Interaction (DTI) is the process by which a drug binds to a specific target site, resulting in a modification of the target's function or behavior [1]. The drug refers to various chemical compounds that, upon consumption and entry into the body, induce physiological changes in the human body. The target refers to living organisms that will bond with the drug and produce physiological changes, which in this case are proteins or nucleic acids [33]. The results of DTI predictions are applied in various drug-related fields such as drug discovery, drug repositioning, drug re purposing, and drug side effects [1, 3]. DTI identification is generally carried out by direct laboratory testing. Establishing drug-target interactions using this experimental approach is both time-consuming and costly [33]. The cost and time required for invitro research in the laboratory are huge, so computational techniques are needed in this field [33]. In silico DTI research using computational techniques and machine learning has been proven to make accurate predictions [1].

Machine learning approaches encounter significant challenges when predicting drug-target interactions, primarily due to dataset characteristics such as class imbalance and the absence of true negative interactions [27, 33]. Class imbalance, where one class overwhelmingly dominates in binary classification tasks, complicates algorithm performance as most models assume balanced class distributions [33]. Consequently, addressing imbalance through preprocessing techniques is crucial for improving predictive accuracy [33]. Strategies to mitigate class imbalance typically involve oversampling, undersampling, or hybrid methods at the data level [15].

Undersampling techniques reduce class imbalance by removing instances from the majority class, which is effective for datasets with moderate class imbalance [9]. Oversampling techniques, such as Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN), aim to tackle significant class imbalance by either duplicating instances from the minority class or creating synthetic ones directly. [29]. Oversampling preserves all original data points, thereby retaining valuable information. This approach helps prevent information loss and maintains the diversity of the majority class, enhancing model generalization and performance [19]. However, oversampling can lead to overfitting, especially when replicating existing samples. Moreover, the process may be computationally intensive, particularly for large, imbalanced datasets[22].

Undersampling techniques can be used with or without optimization algorithms. However, methods like Random Undersampling, which lack optimization, may remove critical information due to their random nature [22]. In contrast, optimization algorithms, though more complex, provide tailored solutions that optimize model performance. They help address class imbalance while preserving dataset integrity and enhancing model efficacy.

This study tackles class imbalance in DTI prediction datasets by using an oversampling technique sequential followed by the undersampling technique. Random Undersampling method and Ant Colony Optimization (ACO) algorithm based undersampling is used to reduce the data that generated by oversampling technique. This approach aims to improve prediction performance by creating a balanced and optimized dataset from a synthetically generated dataset. The minority dataset is first expanded using an oversampling technique, surpassing the number of majority class samples to enhance data diversity. Followed by the use of an undersampling approach to the majority class, a balanced dataset is obtained. The effectiveness of this technique is assessed by evaluating the performance of the generated samples. Experiments with various strategies combining several oversampling existing method and undersampling indicate that integrating optimization algorithms in sampling techniques enhances the classifier's performance in predicting DTI.

1.2 Theoretical Framework

Drug Target Interaction (DTI) is the process by which a drug binds to a selective target site, leading to a modification of the function or behavior of that target [1]. In this context, a drug encompasses various chemical compounds that, once ingested and absorbed into the body, induce physiological changes. Targets, on the other hand, typically refer to biological molecules, particularly proteins or nucleic acids, that interact with the drug to produce these physiological effects. The identification of Drug Target Interactions is essential for understanding the mechanisms of drug action and for the development of new therapeutic agents.

Traditionally, the identification of DTIs has been conducted through direct laboratory testing. This experimental approach involves techniques such as high-throughput screening, biochemical assays, and structural biology methods to observe the binding and effects of drugs on their targets. While this method is effective and can provide detailed insights into the interactions, it is also time-consuming, labor-intensive, and costly [4]. These limitations have led to the exploration of alternative approaches to DTI identification.

In recent years, in silico DTI research, which utilizes computational techniques and machine learning algorithms, has gained prominence. This approach involves the use of computational models to predict DTIs based on various data sources, such as chemical properties of drugs, structural information of targets, and biological activity data. Machine learning techniques, including deep learning and other advanced algorithms, have shown great promise in accurately predicting DTIs. In silico methods offer significant advantages in terms of speed and cost-efficiency, enabling the screening of vast libraries of compounds and the identification of potential drug-target interactions with high accuracy. This computational approach complements traditional experimental methods, providing a powerful tool for drug discovery and development [28].

There is an issue in predicting drug-target interactions using machine learning regarding the available datasets. High complexity and dimensionality are characteristics of the datasets used for DTI prediction. The main problem in predicting DTI is the need for true negative interactions and extreme class imbalance [27]. Class imbalance is the condition in binary classification where one class is represented by a greater number of samples (majority) than another class (minority). The use of imbalanced classes can harm data classification. It can cause the learning classification technique not to achieve good performance with imbalanced data.

Several methods have been proposed to address the imbalanced issue. On the data level there are oversampling and undersampling technique. The oversampling strategy augments the numerical representation of the minority class by either duplicating preexisting cases from the minority class or by creating new ones [17]. Oversampling retains valuable information by preserving all original data points, thus preventing information loss from the dataset. Examples of oversampling techniques include the ROS, SMOTE, and ADASYN.

A commonly employed approach to achieve dataset balance is the SMOTE, which involves the generation of synthetic data points within the minority class. This technique leverages nearest neighbor and interpolation methods to create these synthetic samples [6]. Specifically, SMOTE involves selecting random data points from the minority class and then identifying their nearest neighbors. Synthetic instances are created by interpolating between the points belonging to the minority class and their closest neighbours. This process successfully increases the size of the minority class and improves the model's capacity to learn from these instances. In contrast, to equalize the dataset, the ROS approach employs random replication of instances from the minority class with replacement until the class distributions are balanced [23]. This method does not generate new data but rather increases the representation of existing minority class samples in the dataset. While straightforward, ROS can lead to overfitting because it merely replicates data points rather than providing additional, potentially informative examples. The ADASYN technique enhances the SMOTE approach by focusing on generating synthetic samples based on the difficulty of learning each minority class instance [13]. ADASYN specifically targets those minority class samples that are more challenging for the classifier to learn. By concentrating on these difficult cases, ADASYN aims to improve the classifier's performance on hard-to-learn instances, thereby increasing the overall accuracy and robustness of the model. This method thus provides a more nuanced approach to oversampling by addressing the class imbalance with a focus on enhancing model performance where it is most needed.

Undersampling techniques address class imbalance by selectively removing instances from the majority class, thereby creating a more balanced dataset. This approach is particularly effective for datasets exhibiting moderate class imbalance, where the disparity between the majority and minority classes is significant but not extreme [9]. Random Undersampling (RUS) is a commonly employed algorithm for undersampling. In order to address class imbalance, this method employs a random selection procedure to eliminate instances from the majority class until the number of instances in the majority class matches that of the minority class. The simplicity and ease of implementation of RUS make it an attractive option for balancing datasets. Nevertheless, although RUS is successful in reducing class disparity, it has significant constraints. An inherent limitation of RUS is its indiscriminate elimination of instances from the majority class, therefore resulting in the potential loss of valuable and probably informative data [2]. By discarding a random subset of the majority class instances, RUS may eliminate data points that contain critical information about the underlying distribution of the majority class. This loss of information can adversely affect the performance of the classifier, as the reduced dataset may not fully represent the characteristics of the majority class. Consequently, while RUS helps achieve a balanced class distribution, it risks compromising the quality of the data and the model's ability to generalize effectively.

This work implements the Ant Colony Optimization (ACO) algorithm as a sequential undersampling strategy to be combined with various oversampling techniques. Informed by the natural foraging activity of ants, the ant colony optimization algorithm is an intelligent, heuristic technique designed to improve the selection process for instances to be eliminated. By simulating the pheromone trail-laying and following behavior of ants, the algorithm identifies optimal instances for removal, maintaining the integrity of the dataset while reducing class imbalance. This method is expected to preserve more informative instances from the majority class compared to random undersampling, thereby improving the performance of the classifier.

ACO is a metaheuristic algorithm that draws inspiration from the foraging behavior of actual ant colonies. Successful application of this method has been demonstrated in solving several discrete combinatorial optimization issues [34]. Ants possess the ability to locate the most direct path between their nests and food sources, even in the absence of visual guidance. Additionally, ants demonstrate adaptability to environmental changes and can discover alternative routes when their previous paths are obstructed [24]. ACO in this study operates by removing instances from the majority based on ACO algorithm. The main objective of undersampling using ACO is to find the best instances based on the model evaluation. ACO works by mimicking the way ants leave pheromone trails when foraging. The ants follow these pheromone trails, which guide them towards the optimal path.

1.3 Statement of the Problem

Predicting Drug Target Interactions (DTI) presents significant challenges due to the need for accurate identification of true negative interactions and the issue of extreme class imbalance. In binary classification issues, class imbalance occurs when one class (the majority) is represented by a far greater number of samples than the other class (the minority). The presence of this imbalance can significantly hinder the efficiency of classification methods, since models trained on imbalanced datasets frequently struggle to adequately acquire the distinct features of the minority class. To mitigate class imbalance, various approaches have been suggested, with a main emphasis on oversampling and undersampling methodologies.

Oversampling techniques aim to balance the dataset by increasing the number of instances in the minority class. This can be achieved either through the replication of existing minority class instances or by generating synthetic data points. Oversampling helps retain valuable information by preserving all original data points, thereby minimizing the risk of information loss. In contrast, undersampling techniques address imbalance by selectively removing instances from the majority class to achieve balance. This method is particularly suitable for datasets with moderate class imbalance, where the imbalance is significant but not extreme. However, a notable limitation of undersampling is that it indiscriminately removes instances from the majority class, which can result in the loss of valuable and potentially informative data. The removal of these data points may compromise the dataset's ability to represent the majority class adequately, thereby negatively impacting the performance of the classifier.

The primary problem addressed in this study is the gap in existing balancing techniques for DTI prediction. Specifically, the study aims to evaluate and compare the effectiveness of classic oversampling, sequential hybrid sampling, and sequential hybrid sampling combined with the Ant Colony Optimization (ACO) algorithm. By investigating these methods, the study seeks to determine their impact on improving classification performance and balancing datasets for DTI prediction, ultimately addressing the limitations and gaps present in current balancing techniques.

Based on this condition, statement of the problem that answered in this study are:

1. What is the best classifier on DTI prediction that performed with original dataset?
2. How does the implementation of oversamplig classic technique affect classifier performance in DTI classification?
3. How does the implementation of sequential hybrid sampling using Random Undersampling technique affect classifier performance in DTI classification?
4. How does the implementation of sequential hybrid sampling using ACO based undersampling affect classifier performance in DTI classification?

1.4 Objective and Hypotheses

This study employs a sequential hybrid sampling approach, combining oversampling with undersampling techniques, to address class imbalance in the Drug-Target Interaction (DTI) prediction dataset.The first stage involves applying the oversampling technique to the minority class, which generates synthetic samples to enhance its representation compared to the majority class. This methodology guarantees sufficient representation of the minority class, which is essential in avoiding classifiers from developing bias towards the majority class.

However, relying solely on oversampling may lead to overfitting, as the model might learn to memorize the synthetic instances rather than generalizing from the true underlying data patterns. To mitigate this, the study incorporates an undersampling technique in the second phase, specifically using an Ant Colony Algorithm (ACO). This technique strategically reduces the majority class instances, thereby achieving a more balanced dataset overall. By sequentially applying oversampling followed by undersampling, this hybrid method not only addresses the class imbalance but also enhances the classifier's ability to generalize, leading to improved prediction accuracy in the DTI context.

Based on this hypothesis, this research has the following objectives.

1. To perform baseline classifier evaluation without applied sampling technique to proving the best classifier
2. To applied classic oversampling technique
3. To applied sequential hybrid sampling using Random Undersampling
4. To applied sequential hybrid sampling with ACO

1.5 Assumption

This study is based on the dataset derived from the research presented in [21]. It is assumed that this dataset provides a comprehensive and accurate representation of the features necessary for Drug Target Interaction (DTI) prediction. The correlation of each feature in a classification task can be measured using several statistical methods. However, this study focuses on addressing the issue of class imbalance in the dataset, which contains 680 features. Calculating the correlation for each feature would be computationally intensive; therefore, feature correlations were not measured. Based on this, the following assumptions are made regarding the correlation of features in the dataset:

1. Relevance of Features: The assumption is that all features included in the dataset provide significant and relevant information for the DTI prediction strategy. It is assumed that each feature contributes to the precise representation of the interactions between molecules and their target sites.

2. Feature Integrity: Assuming that all features are pertinent and have been accurately retrieved and processed, these features are expected to contribute efficiently to the prediction of drug-target interactions. The integrity and quality of the features are crucial for ensuring the accuracy and reliability of the study's results.

1.6 Scope and Delimitation

The dataset used in this study is sourced from the study [18]. Their study utilized a comprehensive collection of drug-target interaction information from various databases such as KEGG BRITE and DrugBank. The data collected were drug chemical structures (SMILE format) and protein sequences (FASTA format) from released sources by [32]. The dataset that processed consist of 1404 rows data with 680 features. There are 1334 (93.67%) data with negative label, and 90 (0.06%) data with positive label.

In this study, the independent variables are the features extracted from each drug and target. The drug compounds are represented in the FP2 fingerprint format, a widely-used molecular representation that encodes the presence of particular substructures within the molecules. These features were obtained using OpenBabel software, which provides a robust platform for converting chemical file formats and performing cheminformatics operations. The target features were derived from various descriptors that characterize the amino acid sequence of a protein. These descriptors include methods that assess amino acid composition, which provides information on the frequency of each amino acid within the protein sequence, and dipeptide composition, which examines the frequency of amino acid pairs. Additionally, the descriptors consider amino acid relationships, capturing the sequential or spatial relationships between amino acids, as well as other relevant information extracted from the protein sequences. This comprehensive characterization of the target features ensures a detailed and informative representation of the proteins involved.

The dependent variable in this study is the class, which signifies the predicted interaction between the drug and target. This variable is binary, with Class 0 indicating no interaction (negative interaction) between the drug and target, and Class 1 indicating an interaction (positive interaction) between them. This classification framework allows for the evaluation of the likelihood of interaction, providing valuable insights into potential drug-target interactions and aiding in the identification of promising drug candidates.

This study focuses on the implementation of various sampling techniques to address the challenges posed by imbalanced datasets in classification tasks. Imbalanced datasets, characterized by underrepresentation of particular classes, frequently result in subpar performance of classifiers because of the classifier's reliance on the majority class. To mitigate this issue, the study explores the effectiveness of three specific sampling techniques: oversampling, a sequential hybrid approach combining existing undersampling methods, and a sequential hybrid approach utilizing the Ant Colony Optimization (ACO) method.

The process of oversampling entails augmenting the quantity of instances in the minority class in order to achieve equilibrium within the dataset, whereas undersampling diminishes the amount of instances in the majority class. The objective of the sequential hybrid approach is to combine the benefits of oversampling and undersampling to enhance the classifier's performance. The use of the ACO method in the sequential hybrid approach introduces a novel optimization technique inspired by the foraging behavior of ants, which helps in efficiently finding optimal solutions.

By deliberately not employing feature selection or other data preprocessing methods, the study aims to isolate and understand the direct impact of these sampling techniques on the classifier's performance. This approach ensures that any observed improvements in classification accuracy can be attributed solely to the sampling techniques themselves, providing clear insights into their effectiveness in handling imbalanced datasets.

1.7 Significance of the Study

This study addresses the challenge of class imbalance in Drug Target Interaction (DTI) prediction datasets through the application of a sequential hybrid sampling technique. The sequential hybrid sampling approach involves a two-stage process: first, it applies an oversampling technique to increase the representation of the minority class, followed by an undersampling technique to further refine the dataset. Specifically, this study utilizes an oversampling method in conjunction with the Ant Colony Optimization (ACO) algorithm for undersampling.The sequential hybrid sampling technique begins with oversampling, which generates synthetic instances to augment the minority class and improve its representation in the dataset. Following this, the ACO algorithm is employed for undersampling, which optimizes the selection of instances from the oversampled dataset. This combination aims to achieve a balanced and optimized dataset by leveraging both synthetic data generation and intelligent instance selection.

The significance of this approach lies in its potential to enhance prediction performance by creating a more balanced and informative dataset. The effectiveness of this technique is evaluated by assessing the performance of classifiers trained on the balanced dataset. Preliminary experiments using various strategies that integrate oversampling with undersampling and optimization algorithms indicate that such hybrid approaches can significantly improve the classifier's ability to predict DTI.