CHAPTER 1
INTRODUCTION

1.1 Introduction

Drug discovery is a drug development process that identifies a potential new drug. One of the processes used to identify new drugs is testing the drug's interaction with the target, defined as drug-target interaction (DTI). In this case, a specific protein such as receptors, enzymes, ion channels, etc., is considered the target [2]. The detection of DTI was conducted to achieve a phenotypic effect [2]. Recently, the DTI was identified by laboratory experiments. However, these experiments waste many resources, such as time and cost [2].

To handle this issue, one of the advantages of the computational method is that it can be used as an alternative to predict DTI, such as machine learning [3]. This method can be implemented assuming the drugs have a similar target [4]. One of the challenges in DTI is the significant difference between interacting and non-interacting samples, resulting in an imbalanced dataset that can reduce the performance [4].

To handle this imbalance issue, various balancing methods have been implemented to improve model performance. There are three general sampling approaches, i.e., under-sampling, oversampling, and hybrid-sampling [5]. Under-sampling is a technique where the majority class is reduced by removing instances from the dataset until it is balanced [5]. Oversampling is a technique to generate a new instance as a representation from the minority class until it is balanced [5]. Hybrid sampling is a combined technique where under-sampling is used to reduce the majority class, and oversampling is used to increase the representation of the minority class [6]. In 2019, Ping et al. performed a study about predicting target genes and drugs [1]. They used random sampling using the GradientBoosting Decision-Tree Based Method. The model was evaluated by calculating the area under the curve (AUC) [1]. This study achieved 0.877 on AUC [1]. In 2020, Mahmud et al. developed a drug-target interaction prediction based on a deep learning method [7]. The model of this study implemented a Random sampling with the DeepAction method [7]. AUC calculated the performance evaluation achieved 0.9836 [7].

In 2019, Saad et al. performed a study about predicting drug-target interaction based on adenosine receptors [3]. This study used a combination of four datasets [3]. The study implemented the Synthetic Minority Oversampling Technique (SMOTE) with three classification methods, i.e., Random Forest, Decision Tree, and Support Vector Machine, with a balancing method [3]. The evaluation method used accuracy achieved 75.09% for Random Forest [3]. A study about predicting drug targets based on evolutionary information and chemical structure was performed by Han Shi et al. in 2019 [4]. The four datasets used are enzyme, ion channel, G-protein, and nuclear receptor, which were taken from KEGG BRITE and implemented SMOTE with Random Forest [4]. This study shows that Random Forest is the best method based on the overall accuracies on enzyme, ion channel, G-protein, and nuclear receptor achieved 98.09%, 97.32%, 95.69%, and 94.88 % [4].

Another related study was performed by Zhao et al. in 2021 about identifying drug-target interaction based on graph convolutional networks and deep neural networks [8]. This study shows that GCN-DTI has improved accuracy compared to other methods [8]. A study about drug-target interaction by integrating drug fingerprints and drug side effects using machine

learning was performed by Saad et al. in 2020[9]. This study shows that KNN is the best method in the three experiments, with an accuracy of 96.69% [9]. Another related study about drug-target interaction using a kernel-based framework was performed by Mahmud et al. in 2021 [10]. This study used a dataset from DrugBank [10]. This study shows that the multi-kernel-based method is the best compared to other existing methods, with AUC achieving 0.988 [10]. However, there is no related literature to our proposed method because it is rare.

According to the literature survey, under-sampling is one of the most efficient balancing methods. However, the drawback is the possibility of reducing the potential majority class sample [5]. To overcome this issue, US can be combined with metaheuristic algorithms such as Firefly Algorithm Under-sampling [11]. This algorithm has advantages that can be applied to a wide range of optimization. Hence, this algorithm can improve performance accuracy [11]. Although this algorithm has been used in various optimization processes, its specific application to handle imbalanced data in DTI prediction is innovative.

Based on several related studies, there is an opportunity to develop a new method, Firefly Algorithm Under-sampling. The classification will be implemented using a Random Forest algorithm to measure the impact of the Firefly Algorithm Under-sampling on prediction improvement. Therefore, in this study we aim to handle the imbalanced data on drug target interaction prediction by implementing a Firefly Algorithm Under-sampling. The main focus of this study is to evaluate the effectiveness of the Firefly Algorithm in handling imbalanced data in DTI dataset to improve prediction performance.

1.2 Theoretical Framework

The theoretical framework for handling imbalance issues for drug target interaction prediction can be expanded to the under-sampling principle with the combination of an optimization algorithm, specifically the Firefly Algorithm. This algorithm was developed with the latest innovations to help improve the performance of predictive algorithms.

1. Drug-Target Interaction Theory
   This theory is about an interaction between drugs and their targets, such as enzymes, proteins, etc. The interaction between drugs and targets is an integral part of drug discovery. In pharmacology, DTI is vital for scientists, allowing them to know the drug's mechanism and reduce its side effects.
2. Imbalance Data Theory
   Not all drugs can interact with their intended targets. This occurs because drugs are designed for specific molecular targets such as proteins, enzymes, and receptors. This interaction depends on the drug's molecular structure being compatible with the target. Therefore, it can lead to imbalanced data when the total of non-interacting pairs is more significant than the total number of interacting pairs, reducing the prediction performance.
3. Balancing Method Theory
   The balancing method handles the imbalance of total positive and negative labels. This method can handle the imbalance of total negative and positive and negative labels, reducing the prediction's performance. Many balancing methods exist, such as under-sampling, oversampling, and hybrid sampling.
4. Optimization Algorithm Theory
   An optimization algorithm is a method to find the best solution among sets of

possible solutions for a given problem. In drug-target interaction, this algorithm can improve the efficiency of predictive models in identifying drug candidates.

5. Machine Learning (ML) Theory

   Machine learning, including classification algorithms, is critical in predicting whether a drug interacts with a specific protein. This algorithm can utilize features and chemical properties of drugs and proteins. The algorithm will train on the labelled datasets and evaluated using evaluation metrics. Machine learning can provide more information for the discovery of new drugs

1.3 Conceptual Framework/Paradigm

The main focus of this study on handling the imbalance data of drug target interaction with the implementation of Firefly algorithm under-sampling, the following variables are identified and discussed:

1. Drug Molecules (SMILES)

   Simplified Molecular Input Line-Entry System (SMILES) notation represents a drug chemical molecule that serves as the algorithm's input. Drug chemical molecules have a specific characteristic influencing their interaction with the target (protein).

2. Target (Proteins)

   The target (protein) is the biological proteins within the human body that interact with drug molecules. The interaction between drug and targets can result in adverse reactions and impact a biological response.

3. Predicted drug target interaction

   The predicted drug target interaction has a total difference between interacting and non-interacting pairs, resulting in imbalance issues.

4. Firefly Algorithm Under-sampling

   The Firefly algorithm is one of the optimization algorithms. This algorithm aims to find optimal configurations for under-sampling imbalanced datasets to improve the performance prediction. The algorithm adjusts the sampling strategy based on the natural interaction of virtual fireflies. This method ensures that the dataset is balanced effectively to improve the prediction.

5. Prediction Algorithm (Random Forest Algorithm)

   The Random Forest Algorithm, implemented using the SK-learn framework, is introduced to predict the model. Multiple decision trees developed this model. This algorithm is used to improve predictive performance by preventing overfitting.

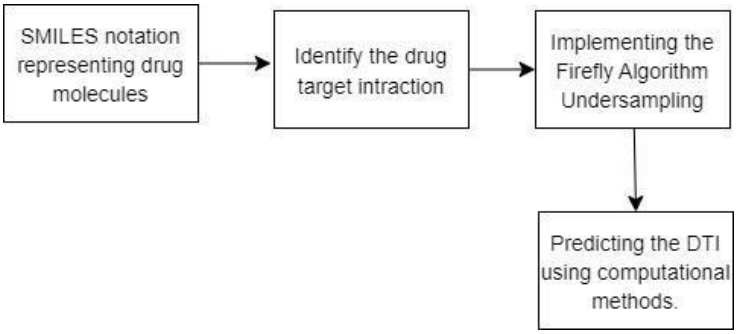

Figure 1 The conceptual framework

1.4 Statement of the Problem

The interaction between drugs and targets varies because not all drugs can interact with all targets. This discrepancy has become a significant issue in drug-target interaction (DTI) research, impacting the performance of prediction algorithms. The issue can be addressed by implementing balancing methods. Implementing the Firefly algorithm under-sampling can handle the imbalance issue in DTI datasets, thereby enhancing the performance of predictive models.

1.5 Objective and Hypotheses

1.5.1 Objective

1. To explore how FAUS influence the balance of class distribution on DTI
2. To examine FAUS's impact on Drug Target Interaction Predictions
3. To Analyze the performance comparison of baseline model performance with and without Firefly Algorithm under-sampling.

1.5.2 Hypotheses

Firefly Algorithm Under-sampling is more effective in handling imbalance data issue on DTI compared to others.

1.6 Assumption

The assumption for implementing the Firefly algorithm under-sampling to handle imbalanced data in drug-target interaction include:

1. The drug-target interaction dataset has poor quality that will affect the prediction performance
2. The DTI dataset has a significant imbalance class ratio.
3. The chosen parameters of Firefly Algorithm will affect the prediction algorithm's performance
4. It is assumed that the Firefly Algorithm can be implemented on larger or complex dataset

1.7 Scope and Delimitation

The main focus of this research is the Firefly algorithm under-sampling to handle the imbalance issue of DTI. The performance of the prediction method will be evaluated using the accuracy, precision, recall, and F1-score.

1.8 Significance of the Study

This research aims to solve the imbalance issue in drug-target interaction Datasets by implementing the Firefly algorithm under-sampling. The implementation of this algorithm enhances prediction performance in drug target interaction research. The Firefly algorithm under-sampling can be applied to complex datasets, making it a valuable method for handling imbalances in real-world data scenarios in drug discovery.