CHAPTER 1
INTRODUCTION

1.1 Rationale

Drugs are chemical substances that engage with a specific target protein, to disrupt biological systems through various molecular interactions [1], [2]. However, drug molecules could interact with non-target proteins, resulting in potential adverse effects. This means that the drug can potentially cause positive or negative changes in the human organism. This condition is known as adverse drug reactions or side effects [1], [3]. Based on data released by the US Food and Drug Administration (FDA), it was found that 90% of experimental drug compounds failed to get FDA approval in clinical trials [4]. This was due to problems with efficacy, formulation, clinical safety, toxicology, and pharmacokinetics (PK) [4]. Of the cases received by the hospital, as many as 5% were ADR cases, and 10%-20% of inpatients who were in the hospital at least once had an ADR case during their stay in the hospital [3], [5]. Due to that concern about the side effects of the drug, there is an immediate need to predict the side effects of the drug [6].

Several approaches are commonly used to predict side effects in drug discovery process, such as in vivo and in vitro [7]. However, the in-vivo method has weaknesses, such as safety procedures requirements, which are relatively low and less accurate [7], [8]. Meanwhile, the in-vitro approach also has the disadvantage of requiring high costs and the inefficient way process [1], [7], [9]. Hence, another approach is needed to handle these weaknesses. An approach that can be used to predict side effects is the in-silico approach. It applies numerical and computational tools to predict drug side effects [3], [10].and also has the advantage in reducing risks in the experimental process by using this approach compared to using other approaches [2], [9].

Several studies related to side effects have been performed, one of them was performed by Saad, A., et al. [11]. They combine 3D information from the ligand and the target [9]. They also used ElectroShape, a method for ultrafast comparison of the validated morphologies and charge distributions of ligands to predict side effects [11]. Other side-effect implementations can be seen in [11]–[13] which use different descriptor variants; there are 2D, 3D, etc. The descriptor itself is a numerical form of the properties of a chemical component [14]. However, the descriptor has an essential process in its use, which performs molecular calculations repeatedly and then proceeds to the next stage [14]. This is one of the disadvantages of using descriptors because the calculation of these descriptors is prone to calculation errors [14].

To overcome this weakness a molecular representation can overcome this weakness by using the Simplified Molecular Input Line Entry System (SMILES), where SMILES is built to represent the chemical language in the machine. This representation has a prominent advantage in reducing calculations and having more efficient computational time [15], [16]. Studies that utilize SMILES was performed in 2017 by Goh, G. B. et al developed the SMILES transform into text to obtain structural information, which can then be used to predict chemical properties [17]. They used the RNN architecture LSTM and GRU as the algorithms chosen to study SMILES features and predict their chemical properties [17]. They compare results from the experiments from each architecture using validation RMSE and AUC, showing that the CNN-GRU architecture effectively predicts chemical properties with extensive data [17]. In this study SMILES is applied as a representation of drug molecules that are converted into vectors, also known as SMILES2Vec, so the drug molecule can get into model.

These days, the use of deep learning methods in the drug development process is now attracting academics. This is because deep learning methods can automatically select features from raw data with high dimensions, which is an important point of difference compared to ordinary machine learning methods [18]. This is also attracting the pharmaceutical industry, and they can use it as a solution to speed up the drug screening process in the drug development process on a large scale [19]. This Long Short-Term Memory (LSTM) architecture also uses for the model research because the data that will enter to the model has a text representation [20]. In LSTM there are 3 main components the input gate, forget gate, and the output gate, where these gates help to determine whether the previously data are forgotten or not [20]. Other advantages in LSTM have a strong ability to extract text information and plays an important role in text classification [20].

Behind the acceptable performance of deep learning, there are several weaknesses, such as high computation time, low performance, and traps at local minima. To overcome this weakness, one way that can be done is to use an optimization algorithm as a useful tuning hyperparameter to achieve better results [21]. Monarch Butterfly Optimization (MBO) is an optimization algorithm that can be used to optimize model experiment, because it has a simple structure, strong resilience, and it can solve global optimization problems [22]. In addition, the MBO algorithm is a new optimization algorithm and has great potential for further development and performance improvement. In the future, this algorithm may take advantage of advances in computer science and mathematics.

This study also used many parameters that had a certain range so that the selection of the best parameters was optimal using optimization algorithms. At this point, there have been no studies that have developed an automated mode using the SMILES2Vec-based LSTM with MBO for predicting drug side effects, based on our research. Therefore, this research aims to implement SMILES2Vec-based LSTM with MBO to obtain the best, and more optimal parameters through several experimental schemes. It includes conducting experiments on the layer arrangement involving the use of convolution layers according to the reference research carried out by Goh et. al [17], [23]. and dense layers as fully connected layers that can combine information with the aim of improving model performance in predicting drug side effects on target organs. With this combination of experiments, it is possible to efficiently explore a large parameter space on a SMILES2Vec-based LSTM architecture for the prediction of pharmacological side effects, particularly for conditions of the Blood and Lymphatic Systems. It is crucial to predict how drugs will affect the blood and lymphatic systems in order to protect patient safety and keep an eye on any potential in drug research or in post-market. Apart from that, based on our search, not many studies have been found that develop LSTM-based SMILES2Vec optimized with MBO.

1.2 Theoretical Framework

This research engages in the drug discovery process, using various approaches. One of these approaches is in-vivo studies, which primarily revolve around using living cells and organisms as the main subjects of investigation [10], [24]. This study often encompasses research involving humans or animals, in which the physiological reactions and relationships within an entire organism [10], [24]. Meanwhile, in-vitro experiments entail the isolation of components from living organisms and their combination within controlled settings like test tubes. Lastly the in-silico method, this method employs numerical and computational techniques to simulate various aspects [10], [24].

In constructing this prediction model, the data that represent the chemical structure of drugs is used. Several types of molecular notation representations exist, including SMILES, the most widely used chemical notation [25]. Other molecular representation is SMARTS, which is an extension of SMILES notation, InChI (International Chemistry Identifier), and InChIKey, which is standardized formulas that are freely usable [25]. For this research, SMILES is used as the representation of chemical molecules. Using SMILES has many advantages, first the notation is easier to read, and can be decoded back into molecular graphs [25]. When transforming SMILES notation from the data into the proposed model, numerous techniques can be employed. One of them is the utilization of molecular descriptors such as fingerprints and MACCS keys [25]. Furthermore, the conversion into a vector form is also possible, a concept referred to as SMILES2Vec, proposed by Goh et al. (2017) [26]. This technique involves the conversion process of SMILES notation into a matrix that can be vectorized.

This study also employed deep learning to construct the model, used the LSTM that has advantage of a high ability to extract textual information, playing a pivotal role in text classification [20]. Furthermore, it also engages in an optimization process aiming at enhancing the performance of the LSTM-based model. The NiaPy framework, which is a Python micro framework is applied to implement the nature-inspired algorithms [27]. Within this framework, there are several algorithms available, including Genetic Algorithm (GA), Particle Swarm Optimization (PSO), etc [27]. For optimization in this study, we use NiaPy framework to implement Monarch Butterfly Optimization (MBO) algorithm. This algorithm falls under the category of swarm intelligence metaheuristic algorithms, drawing inspiration from the migratory behavior of monarch butterflies [28].

1.3 Conceptual Framework/Paradigm

This section thoroughly explains the elements that make up the research challenge. The design is depicted in a schematic diagram to show how each element interacts. The following variables and the relationships are recognized and explored to predict drugs side effects. In Fig. 1 as the representation of the conceptual paradigm between the relationship variable.

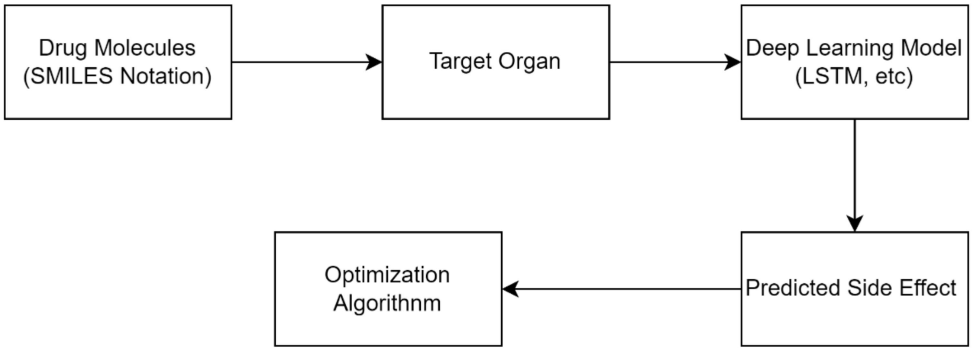
Drug Molecules (SMILES Notation)
Target Organ
Deep Learning Model (LSTM, etc)
Predicted Side Effect
Optimization Algorithnm

Fig. 1 Conceptual Framework Representation

1. Drug Molecules (SMILES) and Target Organs

This is the starting point, representing the chemical structure of the drug molecules in SMILES notation. Drug molecules interact with target organs which are sBlood and Lymphatic System Disorders, and these interactions play a crucial role in determining potential side effects.

2. DL Models (Deep Learning Models) and Target Organ

These models, like LSTM, learn the complex relationships between drug molecules and target proteins to predict potential side effects.

3. DL Models (Deep Learning Models) and Predicted Side Effects

DL models process input data and then, the DL models generate predictions about the potential side effects that could result from the interaction between drug molecules and target proteins.

4. Predicted Side Effects and Optimization Algorithm

An optimization algorithm, like MBO, fine-tunes the parameters of the DL models to enhance prediction accuracy and performance. The optimization algorithm acts as an enhancer, refining the DL model's predictive capabilities.

1.4 Statement of the Problem

Optimization is one way that can be done to improve the performance of the model to be built. This research uses an LSTM architectural model with a combination of convolutional and dense layers, based on previous research conducted by Goh et al in 2018. This study produced very positive results, where the AUC reached a success rate of 88%. This research uses Bayesian Optimization techniques for various types of deep learning architectures, including GRU, LSTM, CNN-GRU, and CNN-LSTM. The research focuses on the context of general chemical property prediction.

In this case study, the optimization process improves model performance. However, the optimization process in the model can also potentially reduce model performance, leading to overfitting of the model. So, this raises the main question, “Does carrying out the optimization process with different approaches and case studies, which is Algorithm MBO and side effect cases on the LSTM model being built increase or decrease the performance of the LSTM model?”.

1.5 Objective and Hypothesis

1.5.1. Objectives

The objectives in this research are:

1. Developing the model architecture with LSTM by SMILES2Vec approach as a representation of the drug-target chemical molecule that has been built as a representation that enters the RNN model used in previous research by Goh et. al [23].
2. Optimizing the model LSTM using the MBO algorithm in searching for optimal parameters for the LSTM model.
3. Evaluating and analyzing the model optimized with MBO using the F1 – Score preformation evaluation metrics and compare the results obtained in the optimized model and the non-optimized model.

1.5.2. Hypothesis

This study builds a model using LSTM with SMILES2Vec representation as in previous research conducted by Goh et. al in 2017-2018 [23], [29]. In this research they developed SMILES2Vec which changed chemical representations in text form in SMILES format into vector representation forms so that they could be processed in RNN models, one of which is LSTM.

Apart from that, MBO is applied, is is as an algorithm that optimizes the LSTM model with SMILES2Vec using the Niapy python framework. This combination is employed with the aim of improving the understanding performance of the model in predicting the presence of side effects on drug-target molecules, namely the blood and lymphatic system.

1.5 Assumption

This study has several assumptions that the interactions between drug molecules and their targets organs (in this study the target organs are the blood and lymphatic system) will determine whether there are side effects. In addition, the chemical representation in this study will be measured using molecular representation in SMILES format, and the SMILES2Vec process helps the SMILES format to enter LSTM model building. Then, we believe the deep learning model we built can understand patterns in the data and use this information to predict the side effects caused by the drug on the target organs of the blood and lymphatic system. Lastly, the assumption in the MBO algorithm optimization is that this algorithm can search for optimal parameters, which can then improve model performance without experiencing overfitting.

1.6 Scope and Delimitation

1.6.1. Principal Variables

This research focuses on two principal variables:

1. Predictive Model: The SMILES2Vec-based LSTM architecture augmented with Monarch Butterfly Optimization which serves as the predictive model. It aims to predict potential drug side effects by leveraging chemical information encoded as SMILES representations.
2. Performance Metrics: The predictive performance is evaluated using F1-Score, precision, and recall assessing the model's ability to accurately predict both positive and negative instances of drug side effects.

1.6.2. Locale

The research is conducted in a computational environment, involving data processing, model implementation, and optimization algorithm execution. Geographical location is not a central consideration, as the focus lies on the algorithmic and computational aspects of the study.

1.6.3. Timeframe

The research is anticipated to span a duration of approximately 12 months. This timeline encompasses data collection, preprocessing, model development, optimization experimentation, and performance evaluation.

1.6.4. Justification

The chosen variables, locale, and timeframe are selected based on several considerations:

1. Research Focus: The central focus is exploring the potential of integrating SMILES2Vec-based LSTM architecture with Monarch Butterfly Optimization to enhance drug side effect prediction accuracy.
2. Scientific Rigor: The scope is designed to maintain a rigorous yet manageable investigation within the stipulated timeframe, ensuring in-depth exploration of the chosen methodology.
3. Applicability: The computational nature of the research allows for broader applicability, with potential relevance to drug development processes across diverse pharmaceutical contexts.

1.6.5. Limitation

Several factors contribute to the delimitations of this research:

1. Data Availability: The research relies on the availability and quality of drug-target interaction data with corresponding side effects, which may influence the model's predictive performance.
2. Model Complexity: This study focuses primarily on the proposed integrated approach of SMILES2Vec-based LSTM architecture and

Monarch Butterfly Optimization, limiting the exploration of other methodologies.

3. Pharmacological Insights: While the model predicts potential side effects, it may not provide detailed mechanistic insights into the underlying pharmacological processes responsible for those effects.
4. Computational Nature: Due to the computational nature of the research, the study does not involve laboratory experimentation or clinical trials.

1.7. Significance of the Study

The contribution of this research is to improve the learning performance of deep learning by using several data transformation techniques, namely SMILES2Vec and the use of the RNN algorithm, namely LSTM, which is referred to in Goh et a.l research in 2017 – 2018 [23], [29]. Apart from that, we also use the MBO algorithm to carry out optimization by carrying out the process search for extensive parameters that fit the model built with the aim of improving model performance and obtaining optimal parameters more quickly. So, through the proposed model it is shown how the performance impact provided by the optimization algorithm has on the results of the evaluation metrics that we use when compared with models that were not optimized using MBO.