Implementation of Camel Algorithm-Support Vector Machine in Predicting Angiotensin Converting Enzyme (ACE) Inhibitor as Antihypertensive Agent

1st Raihan Fathul Bayan
School of Computing
Telkom University
Bandung, Indonesia
raihanfathulb@student.telkomuniversity.ac.id

2nd Isman Kurniawan
School of Computing
Telkom University
Bandung, Indonesia
ismankrn@telkomuniversity.ac.id

***Abstract*—Hypertension is a serious medical condition and a leading cause of death worldwide. The development of accurate predictive methods for Angiotensin Converting Enzyme (ACE) Inhibitors is crucial in the treatment of hypertension. This study aims to develop a predictive model using Camel Algorithm for feature selection and Support Vector Machine (SVM) for building the predictive model with three types of kernels including Radial Basis Function (RBF), SVM-Linear, and Polynomial. The dataset used consists of 255 compounds with ACE inhibitory activity in Rattus norvegicus rats, sourced from the ChEMBL database, structured in SMILES notation, and then converted into SDF format. Experimental results show that the RBF kernel model provides the best performance, with an R^2 value of 0.8728 on the training data and 0.5620 on the testing data. This study highlights the effectiveness of the Camel Algorithm-Support Vector Machine combination in developing the ACE Inhibitor prediction methodology, and provides a significant contribution to the development of safer and more effective antihypertensive drugs, which can pave the way for further advancements in this field.**

Keywords—Hypertension, ACE Inhibitor, Support Vector Machine, Camel Algorithm, Feature Selection

I. INTRODUCTION

Hypertension is a condition where blood pressure in the blood vessels increases consistently [1]. This condition comes without obvious symptoms and difficult to detect without specialized tools [2]. Hypertension is one of the causes of death in the world by contributing to 10.4 million deaths annually in the world [2]. By 2020, more than 1.13 billion people worldwide suffer from hypertension [2]. The main causative factors of hypertension can be divided into fixed (e.g., age, genetics, gender) and non-fixed factors (e.g., lifestyle, body weight, stress levels, and excessive salt intake) [3], [4]. The diagnosis of hypertension is characterized by systolic blood pressure of more than 140mmHg and diastolic blood pressure of more than 90mmHg [4]. This condition has the effect of increasing the risk of various serious diseases such as heart disease, stroke, kidney failure, and others [1]. Therefore, effective treatment is needed to prevent the risk of hypertension and its complications.

In the context of the causes of hypertension, it is crucial to understand the role of Angiotensin-Converting Enzyme (ACE), which is involved in the renin-angiotensin system (RAS) of the human body. ACE's primary function is to convert angiotensin I into angiotensin II, a substance that causes blood vessels to constrict, thereby increasing blood pressure [5]. The conversion of angiotensin I into angiotensin II plays an important role in regulating blood pressure and fluid balance in the body [6]. Overproduction of angiotensin II, which causes blood vessels to constrict, can lead to increased blood pressure and potentially cause hypertension [7]. Therefore, inhibiting ACE activity is an effective way to control blood pressure and prevent hypertension [6]. Therefore, inhibiting the activity of ACE is one of the most effective way to avoid an increase in blood pressure or hypertension [6].

In controlling blood pressure, ACE inhibitors have been shown to be effective by inhibiting excessive ACE enzyme activity, thereby preventing excessive angiotensin II production [8], [9]. The conventional method of developing ACE inhibitors is done through wetlab research involving the separation, screening, and identification of compounds from natural materials [10], [11]. However, this method is time-consuming and costly as it involves complex processes, such as sample preparation, analysis, and the use of laboratory equipment [10]. As an alternative, in silico methods with Machine Learning approaches enable large-scale data analysis quickly and efficiently, making it a more practical solution.

In recent years, there have been several studies predicting the activity of ACE Inhibitors with various Machine Learning methods. In 2019, Elaziz et al. showed that the RVFL method enhanced with the Sine-Cosine algorithm (SCA) had the best performance with an RMSE of 0.12 [12]. In 2020, Wang et al. compared four Machine Learning methods and found that Random Forest (RF) and Support Vector Machine (SVM) had the best performance with R^2 score between 0.29 to 0.51 [13]. In the same year, S K et al. used the Genetic Algorithm-Multiple Linear Regression (GA-MLR) method for the QSAR model with the result of R^2 of 0.66 [14]. In 2021, Wang et al. identified antihypertensive peptides using the XGBoost model, which achieved 94.11% accuracy [15]. Finally, in 2023, Tianshi Yu et al. found that the RF algorithm was superior overall with an accuracy of 0.981, while the Multi-layer Perceptron (MLP) algorithm was better on certain subsets with an accuracy of 0.973 [16]. Based on the reviewed literature, studies predicting ACE inhibitor activity have shown significant progress with various machine learning methods. Despite the good results, the integration of feature selection techniques in study remains underexplored,

which limits further improvements in model accuracy and interpretability.

This study aims to address this gap by integrating Camel Algorithm and Support Vector Machine in predicting ACE Inhibitors as antihypertensive agents. This approach utilizes the ability of Camel Algorithm in selecting relevant features with the power of SVM in building accurate prediction models. This study aims to make a meaningful contribution in improving the effectiveness of ACE Inhibitor prediction through a more careful and optimized feature selection technique.

II. Material and Methods

A. Dataset

The dataset used in this study was sourced from the chEMBL database and prepared using SMILES notation, which was subsequently converted into SDF format [17]. The dataset consists of 255 compounds with ACE inhibitory activity against Rattus Norvegicus rats [17]. Initially, the activity data was expressed in IC50 (nM) and converted to molar units for consistency [17]. The IC50 values were then transformed into pIC50 (-log10IC50) prior to conducting QSAR analysis. The molecular structures of the compounds were optimized using the MMFF94 force field in Marvin by ChemAxon [17]. Before processing the dataset, a filtering step was performed to remove irrelevant or duplicate data. The dataset was then split into training and test sets with a ratio of 70 (train) : 30 (test).

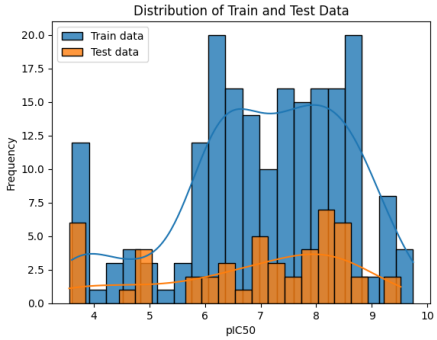

Fig. 1. Distribution of Dataset

In Fig. 1 shows the distribution of datasets for training and test data based on pIC50 values against SMILES. This distribution shows a diverse range of activity values, ensuring adequate representation for the model. This dataset distribution is designed to balance between model training and evaluation.

B. Feature Selection

Camel Algorithm (CA) is a metaheuristic optimization algorithm inspired by the behavior of camels walking in the desert. CA adapts the concept of camel trips (or "kernels" in the context of the algorithm) where each kernel represents a route to be traveled [18]. The camels move in search of the best solution considering vehicle capacity constraints and time windows [18]. There are three main factors in the Camel Algorithm, namely temperature (T), supply (S) and endurance (E). The three factors can be seen in equation below [18].

$$T_{now}^{j} = (Tmax - Tmin)Rand + T_{min} \quad (1)$$

$$S_{now}^{j} = S_{past}^{j} * \left(1 - \omega \frac{Traveled\ stpes}{total\ journey\ steps}\right) \quad (2)$$

$$S_{past}^{j} = S_{now}^{j} \quad (3)$$

$$E_{now}^{j} = E_{past}^{j} * \left(1 - \frac{T_{now}^{j}}{Tmax}\right) * \left(1 - \frac{Traveled\ stpes}{total\ journey\ steps}\right) \quad (4)$$

$$E_{past}^{j} = E_{now}^{j} \quad (5)$$

$Tmax$ And $Tmin$ represent the maximum and minimum temperatures, respectively, which can be selected as needed [18]. At the start of the journey, E_{past}^{j} is initialized to $E_{initial}^{j}$, which represents the maximum endurance. As the journey progresses, the endurance is updated at each step according to equation (5) [18].

CA is used to identify and retain the most relevant features from a given data set by removing uninformative and redundant data [19]. In feature selection, the optimized function is the regression accuracy, which is determined by the number of features selected. This implies that there is a function to be minimized, as shown in equation (6) [19].

$$f(x) = \alpha \times (1 - P) + (1 - \alpha) \times \frac{N_{selected}}{N_{features}} \quad (6)$$

Where α is a parameter that determines the trade off between the performance of the regressor P and the number of features selected [19]. Using CA, to find the optimal or sub-optimal solution to the feature selection problem, the high dimensionality of the features can be overcome by utilizing the parameters described in Table I [18].

TABLE I.
PARAMETER FOR CAMEL ALGORITHM

Parameter	Value
Population size	[15,20,25]
Burden factor	0.25, 0.5
Supply init	[15,20,40]
Endurance init	10, 15

C. Model Development

Support Vector Machine (SVM) is a popular Machine Learning method commonly used for classification and regression tasks. The working principle of SVM is to find the best hyperplane that can separate two classes of data with the maximum margin [20]. A hyperplane is a dimension that divides or classifies data into two different classes [21]. In this algorithm, the data objects that are closest to the hyperplane are defined as support vectors, which play a crucial role in finding the optimal hyperplane. SVM has the advantage of handling complex and unstructured data. Additionally, the performance of SVM can be enhanced through the use of parameter optimization techniques such as Particle Swarm Optimization (PSO), which helps in finding the best hyperparameters to improve the quality of predictions with an accurate and reliable model [22].