A Hybrid ROS-SVM Model for Detecting Target Multiple Drug Types

Nur Ghanaviyanto Ramadhan a,*, Azka Khoirunnisa b, Kurnianingsih c, Takako Hashimoto d

a Institut Teknologi Telkom Purwokerto, Indonesia,
b Telkom University, Bandung, Indonesia,
c Politeknik Negeri Semarang, Indonesia,
d Chiba University of Commerce, Japan

* Corresponding author: ghani@ittelkom-pwt.ac.id

Abstract—Misleading in determining the decision to use the target drug will be fatal, even to death. This study examines five pharmacological targets designated as types A, B, C, X, and Y. Early detection of misleading drug targeting will reduce the risk of death. This study aims to develop hybrid random oversampling techniques (ROS) and support vector machine (SVM) methods. The use of the oversampling technique in this study aims to balance classes in the dataset; due to the data collection in each class, there is a relatively large gap. This study applies five schemes to see which combination of models produces the highest accuracy. This study also uses five types of SVM kernels, linear, polynomial, gaussian, RBF, and sigmoid, combined with the ROS oversampling technique. Our proposed model combines the ROS oversampling technique with a linear SVM kernel. We evaluated the proposed model and resulted in an accuracy of 97% and compared it with several experiments, including the ROS technique with a sigmoid kernel which only resulted in 50% accuracy. It can be seen from the results obtained that the linear kernel is very adaptive to data types in the form of numeric and nominal compared to other kernels. The method proposed in this study can be applied to other medical problems. Future research can be carried out using a combination of other sampling techniques with deep learning-based methods on this issue.

Keywords—Drug; random oversampling; support vector machine; balancing data.

I. INTRODUCTION

A drug is a substance or mixture of materials used to determine, prevent, reduce, kill, cure infection or symptoms of disease, injury, or another physical or worldly disease in humans or creatures’ body [1]. Drugs come in wide varieties, including generics, proprietary generics, and brand-name generics [1]. When patients do not take appropriately targeted medications, regardless of the type of disease, complications develop. For instance, a person with a mental disease who needs medication. If the drug kind is unknown for the illness, it is risky. This issue is hazardous because it can kill those who consume it. Using intelligent machine learning approaches based on machine algorithms to identify which drug to provide can fix the problem.

Problems related to the detection of drug targets have often been carried out by previous researchers [2]-[24] by using several detection models based on deep learning such as Bidirectional-Long Short-Term Memory (BI-LSTM) and Artificial Neural Network [7]-[10], [12]-[20]. In addition to deep learning-based models, some discuss the use of traditional machine learning models such as decision trees, naive bayes, K-Nearest Neighbors [11], random forest [21], and Xtreme Gradient Boosting [23]. However, from several previous studies, few focus on preprocessing data related to data balancing [24].

Nascimento et al., have undertaken numerous experiments on targeting drug users, including the prediction of drug-target interaction problems using the KronRLS-MKL model [2]. Yasuo et al. proposed a study to deal with the prediction of drug-target interactions, the challenge was to recognize new protein-ligand interactions from prior information based on deep learning [3]. Olayan, et al developed efficient computational method, a modern strategy that allows drug–target interactions (DTI) to measure the accuracy of expectations. efficient computational method is based on the use of heterogeneous graphs, with drug–target interactions findings having much in common between drugs and little in common between target proteins [4].
Ryu et al. The accuracy of the deep learning method in predicting drug-target and drug-food component interactions was 92% [5]. The same problem uses an artificial neural network (ANN) model [6]. Other experts offer the Drug-Target Intelligent Bayesian Positioning Forecast (BRDTI). This concept is based on the factorization of the Bayesian Personalized Positioning (BPR) network, which has proven to be an effective technique for analyzing various trends [7]. This model has not previously been used for drug–target interactions prediction [7].

Ezzat Ali et al. predict a previously unforeseen relationship between class imbalance and another unresolved issue with the potential to impair predictive performance [8]. Bagherian et al. focus on predicting drug-target interactions by playing a role in drug discovery [9]. Chen Roulan et al. Focus on machine learning techniques and give a complete understanding of drug-target interaction prediction [10]. Lavecchia et al. discuss machine learning techniques used to predict drug-target interactions, such as supporting vector machines, decision trees, naive Bayesian, K-Nearest Neighbors, and ANN [11]. Liu Yong et al. introduced a technique referred to as regular logistic matrix factorization for predicting drug-target interactions (NRLMF) [12]. Peon, et al reported the prediction of drug-target interactions using the maximum coverage approach [13]. Mohammed Nazim, et al Present the SimBoost approach for continuous (non-binary) prediction of drug-target interactions [14]. The Bi-LSTM model was studied to predict interaction problems between drugs [15].

Wongyikul et al. regarding medication sentiment analysis, which has become crucial for classifying existing drugs according to their efficacy. The investigation was conducted using user reviews that assist possible future consumers in gaining knowledge and making more informed selections about a specific drug [16]. Shanbhag et al. developed a screening protocol for high alertness medication (HAD) in 2018, high alertness drug prescribing (HAD) errors in inpatient and outpatient prescribing at Maharaj Nakhon Hospital in Chiang Mai were identified utilizing a machine learning model using Gradient Boosting Classifier and screening parameters [17].

Liu Bin et al. discuss drug target interaction using the Nearest Neighbor weighting technique and sampling the drug probabilities. However, in this study, it is not explained what the sampling method is only local sampling is mentioned [18]. Thafar Maha et al. discuss drug prediction using a graph-shaped approach and name similarity. In this study, the number of each class in the dataset is not checked [19]. Pliakos et al. in this study predict drug targets using a tree-based machine learning model approach. The tree model here is still done conventionally [20]. Mohan et al. compare ensemble learning models, single models, extra tree, and random forests to predict drug targets. This study does not discuss the imbalanced technique in the dataset used [21].

Ye Qing et al. used a multiple output deep neural network approach to predict drug targets, which resulted in multiple layers, but this study did not discuss preprocessing data [22]. Mahmud et al. demonstrated the use of data balancing techniques for the prediction of drug targets [23]. The study also uses a classification model other than deep learning, namely Xtreme Gradient Boosting; the dataset used has four classes [23]. Mahmud et al. aim to predict drug targets by applying the SMOTE oversampling technique and the Xtreme Gradient Boosting algorithm as predictions [24].

Based on the explanation of the problems above, this study aims to detect several types of drugs by applying a dataset balancing technique with a detection model using a support vector machine model approach.

The main contributions to this research are:
1. This research implements an oversampling technique called random oversampling (ROS) to equalize the amount of classes inside a dataset.
2. This research implements a machine learning support vector machine (SVM) model with five kernels: linear, polynomial, gaussian, RBF, and sigmoid.

The advantage of using the support vector machine method that is hybridized with random oversampling techniques will make the data more accurate in detecting or classifying. We evaluate the proposed model using accuracy and correlation matrix. The structure of this paper consists of Section 2 consists of research flow diagrams, data collection, and methods. Sections 3 and 4 discuss the proposed methods application, results in analysis, and conclusions.

<table>
<thead>
<tr>
<th>Author</th>
<th>Balancing Dataset</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ryu et al [7]</td>
<td>No</td>
<td>Deep learning</td>
</tr>
<tr>
<td>Shtar et al [6]</td>
<td>No</td>
<td>ANN</td>
</tr>
<tr>
<td>Lavecchia et al [11]</td>
<td>No</td>
<td>KNN</td>
</tr>
<tr>
<td>Liu et al [18]</td>
<td>Yes</td>
<td>KNN</td>
</tr>
<tr>
<td>Mahmud et al [24]</td>
<td>Yes</td>
<td>XGBoost</td>
</tr>
<tr>
<td>This Proposed</td>
<td>Yes</td>
<td>SVM</td>
</tr>
</tbody>
</table>

II. MATERIAL AND METHODS

This study using the proposed methods can be seen in Figure 1.

![Proposed Diagram](image-url)


A. Data Collection

The data collection used in this study has five types of drugs to be detected with a total of 200 data. The study data was obtained from U.S. market international data. Details of the attributes and dataset types can be seen in Table II.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Type Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Int</td>
</tr>
<tr>
<td>Sex</td>
<td>String</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>Int</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>String</td>
</tr>
<tr>
<td>Potassium Ration</td>
<td>Int</td>
</tr>
<tr>
<td>Drug Type</td>
<td>String</td>
</tr>
</tbody>
</table>

In Table III about distribution dataset in this research.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>44.31</td>
<td>13</td>
<td>74</td>
<td>16.47</td>
</tr>
<tr>
<td>Sex</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>Blood Pressure</td>
<td>1.02</td>
<td>0</td>
<td>2</td>
<td>0.78</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>Potassium Ration</td>
<td>16.08</td>
<td>6.2</td>
<td>38.24</td>
<td>7.19</td>
</tr>
</tbody>
</table>

B. Pre-processing

In this process, checking whether there are rows of data that have no value or are null. If there is, deletion of the data row will be carried out. Furthermore, checking whether there are special characters such as ‘?,”!’ ‘/’, the removal of the character will be carried out. After that, for attributes with string data types, the conversion will be carried out to the integer form using the if-else form. So that the final data that will be used for the process of oversampling and data type detection is in the form of an entire integer.

C. Oversampling Drug Type

Applying a resampling strategy in preprocessing, data processing to get a greater balanced distribution of records is an effective strategy to the imbalance hassle [25],[26]. In random oversampling, the data are supplemented with a random sample of the minority class sample. This can increase the likelihood of overfitting, particularly at greater oversampling values. In addition, it reduces classifier performance and increases computational effort [27]. Additionally, the ROS technique requires randomly replicating a sample of the minority class and including it in the training dataset [26]. This technique has also been frequently used in several health-related issues such as drug targets [24] and diabetes detection [26].

This oversampling process is carried out on the drug target which has a minor class. Figure 3 shows the total distribution of the original data. While in Figure 4 is the result of the distribution of the amount of data after the application of random oversampling (ROS) because this method on the problem of balancing health topics has proven to be effective compared to other oversamping methods [26]. This oversampling technique can simply be seen in Figure 2. In figure 2, the under-sampling technique works by reducing the amount of majority data in the original data, while the oversampling technique works by adding the amount of data to the minority class.

In figure 2, the under-sampling technique works by reducing the amount of majority data in the original data, while the oversampling technique works by adding the amount of data to the minority class.

Figure 3 is the number of datasets before the ROS oversampling technique was applied, where it was seen that the difference in the number of each type of drug was very prominently different.
balanced because it will affect the accuracy of the results that will be obtained. So that after the application of ROS, the amount of data on drug target types A, type B, and type C increased to match the amount of data on type X. Different things happened to drug target type X, which did not experience an increase in the amount of data, this happened because the system assumed the number of data type X is quite a lot. Type B and type C previously had only 16 data, and type A had only 23 data.

D. Classification

Support vector machine (SVM) is a tiny, innovative sample-based learning technique based on the structural risk reduction concept as opposed to the conventional empirical risk minimization theory [28]. It is better than the current method; many see that the Support vector machine is the best 2D illustration of a linear surface formed from separate cases. The basic idea can be used in Figure 5. The two types are separated by H without error. \( H_1 \) \( H_2 \) is the area that passes through the last point H. The distance between \( H_1 \) and \( H_2 \) is called the class interval. It is best to separate the face not only to avoid the error of separating the two types of use, but it is also called the most significant class interval. The SVM method in this study was chosen because, in several studies, it successfully detected various problems, for example, in a problem that discusses the classification of malaria where the data contains parameters related to the disease [29]. In other studies, SVM is used to analyze the results of sentiment toward film reviews [30].

\[
K(x, y) = x^T y
\]

\[
K(x, x_j) = (x^T x_j + 1)^d
\]

\[
K(x, y) = \exp(-\frac{|x-y|^2}{2\alpha^2})
\]

\[
K(x, x_j) = \exp(-r|x-x_j|^2), \quad r>0
\]

\[
K(x, y) = \tanh(ax^T y + c), \quad r>0
\]

Where \( x, y, z, x_i, x_j \) represent the data used. \( x^T \) is vector, \( c \) is bias, \( r \) is degree, and \( \alpha \) is large value.

This SVM method has a name for the kernel. This kernel serves to determine the results and data to be used. SVM has five types of famous kernels, namely linear kernels (2) are suitable to be used with facts with many features, which include textual content facts. Kernel functions and parameters utilized in SVM assessment have a substantial impact on the resulting accuracy [29]. In addition, there is also a polynomial kernel (3) whose function is a non-linear kernel which is very suitable for problems whose all training datasets are normalized. Gaussian Kernel (4) was used whilst there may be no earlier know-how approximately the facts. The Radial Basis Function (RBF) (5) kernel is a kernel function that is used when the data cannot be separated linearly, where in analyzing with RBF, cost and gamma parameters will be optimized. Sigmoid kernel (6) functions more of a model that is deeply defined as neural networks. This study uses linear kernels because the study [29] has succeeded in increasing accuracy.

III. RESULTS AND DISCUSSION

This research will be carried out several experiments to find out what kind of model can produce the highest accuracy and precision.

A. First Schema

This first scheme is done by using a linear SVM kernel and ROS oversampling technique. The results obtained can be seen in Table 4.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Model (SVM linear+ROS)</td>
<td>97</td>
<td>96</td>
</tr>
<tr>
<td>SVM Linear</td>
<td>78</td>
<td>74</td>
</tr>
</tbody>
</table>

Based on Table 4, the accuracy of the proposed model is 97% while the accuracy of SVM Linear is 78%. This shows that the proposed model gives higher accuracy. The results in terms of the precision of this research model proposal are also still high, which is 96% compared to the SVM of ordinary linear kernels.

B. Second Schema

This second scheme is done by using the polynomial SVM kernel and the ROS oversampling technique. The results obtained can be seen in Table 5.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Polynomial+ROS</td>
<td>71</td>
<td>65</td>
</tr>
<tr>
<td>SVM Polynomial</td>
<td>66</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 5 shows that the accuracy of our proposed model is higher than SVM Polynomial, with an accuracy of 71%. The precision results in the second experiment were very different from the accuracy results obtained from whether from the application of oversampling or not. The difference can reach 6%, which means that for the score of the confusion matrix, the false positive value has a high score.
Of course, this is very dangerous because the results are wrong but predicted to be correct.

C. Third Schema

This third scheme is done by using the Gaussian SVM kernel and the ROS oversampling technique. The results obtained can be seen in Table 6.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Gaussian + ROS</td>
<td>76</td>
<td>71</td>
</tr>
<tr>
<td>SVM Gaussian</td>
<td>70</td>
<td>67</td>
</tr>
</tbody>
</table>

Based on Table 6, our proposed model achieved higher accuracy than SVM Gaussian, with an accuracy of 76%. In this experiment, it was better than the second experiment where the difference between precision results and accuracy was only 5%. The difference of 1% in predictions certainly has a great influence on users later.

D. Fourth Schema

This fourth scheme is done by using the SVM Radial Basis Function kernel and the ROS oversampling technique. The results obtained can be seen in Table 7.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM RBF + ROS</td>
<td>77</td>
<td>70</td>
</tr>
<tr>
<td>SVM RBF</td>
<td>70</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 7 shows that our proposed method yielded an accuracy of 77%. It is higher than the SVM RBF model which only gives an accuracy of 70%. The precision result for the application of the oversampling technique is quite far, which is 7% compared to without applying the oversampling technique, which is 9%. In this fourth experiment, of course, that the RBF kernel has a significant decrease in precision value.

E. Fifth Schema

This fifth scheme that is carried out is using a sigmoid SVM kernel and a ROS oversampling technique. The results obtained can be seen in Table 8.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM Sigmoid + ROS</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>SVM Sigmoid</td>
<td>56</td>
<td>51</td>
</tr>
</tbody>
</table>

Based on Table 8, the accuracy of the proposed model is 50%. This accuracy is lower than the accuracy of SVM Sigmoid, which achieved an accuracy of 56%. In this fifth experiment, the application of the sigmoid kernel had the ugliest precision results compared to other kernels. The precision results can only be 40% which is 10% difference after the application of oversampling. Another thing without using the application of oversampling the precision result only dropped by 5%. In this kernel, the oversampling technique is not suitable for use because the results have decreased.

Figure 6 and 7 is a summary of five schemes that have been carried out specifically for the SVM model. Included
the summary of the five schemes by applying the oversampling technique of ROS and various SVM kernels.

Based on Figure 6 it can be analyzed that the highest accuracy results when applying the ROS technique with a combination of SVM linear kernels can produce an accuracy of 97%. The combination of these two techniques is the initial goal proposed in this study. Compared to when only applying the SVM model with a linear kernel, it is only able to produce 78% accuracy; the difference in accuracy reaching 19% is very risky if applied later because it can be fatal for the user and is not in line with the purpose of this study, which is to help determine the target drug used. The user will consume it.

When combined with the ROS technique, the polynomial SVM kernel produces an accuracy of 71%, compared to only using the SVM polynomial model, which is only 66%, the difference in accuracy in the polynomial kernel is 5%. When combined with the ROS technique, the gaussian SVM kernel produces 76% accuracy, compared to only using the gaussian SVM model of only 70%, the difference in this kernel is 6%. In the RBF kernel combined with the ROS technique, the accuracy is 77%, compared to without the ROS technique combination, which is only 70%, and the difference in this kernel is 6%. Another SVM kernel is sigmoid; when combined with the ROS technique, the accuracy is 50%, compared to 56% without the combination of the ROS technique. An exciting thing is seen in the sigmoid kernel is incompatible with the data balancing method application. The linear kernel in this study produced the highest accuracy because the dataset used can be separated linearly, so this research dataset is suitable using a linear kernel type.

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Polynomial</th>
<th>Gaussian</th>
<th>RBF</th>
<th>Sigmoid</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>97%</td>
<td>71%</td>
<td>76%</td>
<td>77%</td>
<td>50%</td>
<td>76%</td>
</tr>
</tbody>
</table>

![Fig. 8 Correlation Matrix Heatmap](image)

Figure 8 is a correlation matrix between variables in the dataset. This research uses heatmap correlation in python for describe feature correlation. This matrix correlation illustrates that if the value is close to 1, the association between the two variables is strong and positive. If the value is outside of 1 or near -1, the link between the two variables is negative or weak. Judging from the correlation matrix in this research dataset, the variables that strongly relate to the drug target are age 0.28 and potassium ratio 0.38.

IV. CONCLUSIONS

In this study, the goal is to predict the drug of multiple class targets using data balancing techniques. After experimentation, then this study succeeded in answering the research objectives, namely the prediction of several drug targets (A type, B type, C type, X type, and Y type). The oversampling technique used, namely random oversampling (ROS), also successfully balanced the amount of class data on each type of drug target. This study also applies five kernels to the support vector machine method: linear, polynomial, gaussian, RBF, and sigmoid. In this study, a combination of oversampling techniques with kernels on SVM is carried out. The results obtained are linear kernels with a combination of ROS techniques get the highest accuracy of 97%, compared to only using the SVM linear kernel only 78%. The second highest kernel accuracy when oversampling is combined RBF at 77%, the three gaussian kernels are 76%, the four polynomial kernels are 71%, and the sigmoid kernel is 50%. Among the five kernels, the sigmoid kernel produces an accuracy of only 50%, even though it has been combined with oversampling techniques.

Further research can be discussed related to the sigmoid kernel on the application of oversampling and undersampling techniques.

ACKNOWLEDGMENT

Authors says thank to Institut Teknologi Telkom Purwokerto have supported this research.

REFERENCES


