INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

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

Nur Ghaniaviyanto 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.**

Manuscript received 3 Aug. 2022; revised 29 Sep. 2022; accepted 31 Oct. 2022. Date of publication 10 Sep. 2023.

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, naïve 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 I
COMPARISON OF PREVIOUS STUDY

Author	Balancing Dataset	Model
Ryu et al [7]	No	Deep learning
Shtar et al [6]	No	ANN
Lavecchia et al [11]	No	KNN
Liu et al [18]	Yes	KNN
Mahmud et al [24]	Yes	XGBoost
This Proposed	Yes	SVM

II. MATERIAL AND METHODS

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

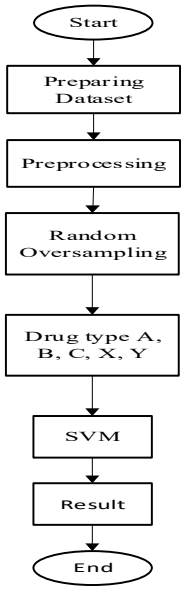

Fig. 1 Proposed Diagram

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 II
DATASET

Attributes	Type Data
Age	Int
Sex	String
Blood Pressure	Int
Cholesterol	String
Potassium Ration	Int
Drug Type	String

In table III about distribution dataset in this research.

TABLE III
DISTRIBUTION DATA

Attributes	Mean	Min	Max	Standard Deviation
Age	44.31	15	74	16.47
Sex	0.48	0	1	0.49
Blood Pressure	1.02	0	2	0.78
Cholesterol	0.5	0	1	0.49
Potassium Ration	16.08	6.2	38.24	7.19

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.

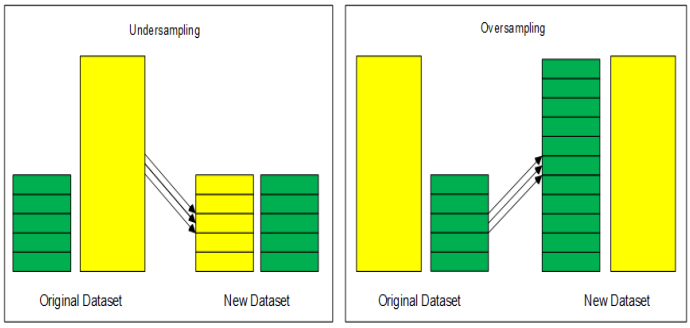

Fig. 2 Oversampling and Undersampling

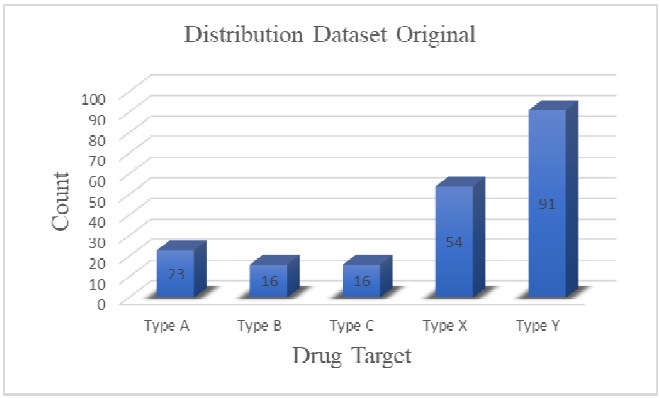

Fig. 3 Before Apply ROS

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.

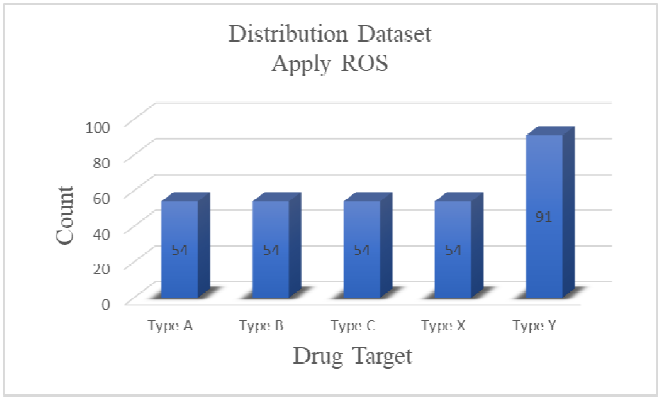

Fig. 4 Applying ROS

Applying this oversampling technique aims to balance the amount of data in each class. This technique is essential in the case of datasets where the number of class data is not

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].

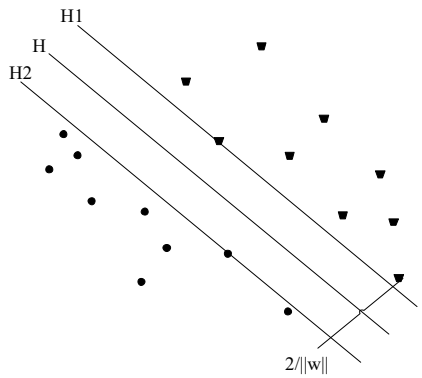

Fig. 5 Hyperplane SVM

$$K(x,z) = x^T z \quad (2)$$

$$K(xi,xj) = (xi.xj+1)^d \quad (3)$$

$$K(x,y) = exp(-||x-y||^2 / 2a^2) \quad (4)$$

$$K(xi,xj) = exp(-r||xi-xj||^2), \ r>0 \quad (5)$$

$$K(x,y) = tanh(ax^T y+c), \ r>0 \quad (6)$$

Where x, y, z, xi, xj represent the data used. x^T is vector, c is bias, r is degree, and a 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 IV
RESULT OF LINEAR KERNEL SCHEMA

Models	Accuracy (%)	Precision (%)
Proposed Model (SVM linear+ROS)	97	96
SVM Linear	78	74

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 V
RESULT OF POLYNOMIAL KERNEL SCHEMA

Models	Accuracy (%)	Precision (%)
SVM Polynomial+ROS	71	65
SVM Polynomial	66	62

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 VI
RESULT OF GAUSSIAN KERNEL SCHEMA

Models	Accuracy (%)	Precision (%)
SVM Gaussian+ROS	76	71
SVM Gaussian	70	67

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 VII
RESULT OF RBF KERNEL SCHEMA

Models	Accuracy (%)	Precision (%)
SVM RBF+ROS	77	70
SVM RBF	70	61

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 VIII
RESULT OF SIGMOID KERNEL SCHEMA

Models	Accuracy (%)	Precision (%)
SVM Sigmoid+ROS	50	40
SVM Sigmoid	56	51

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.

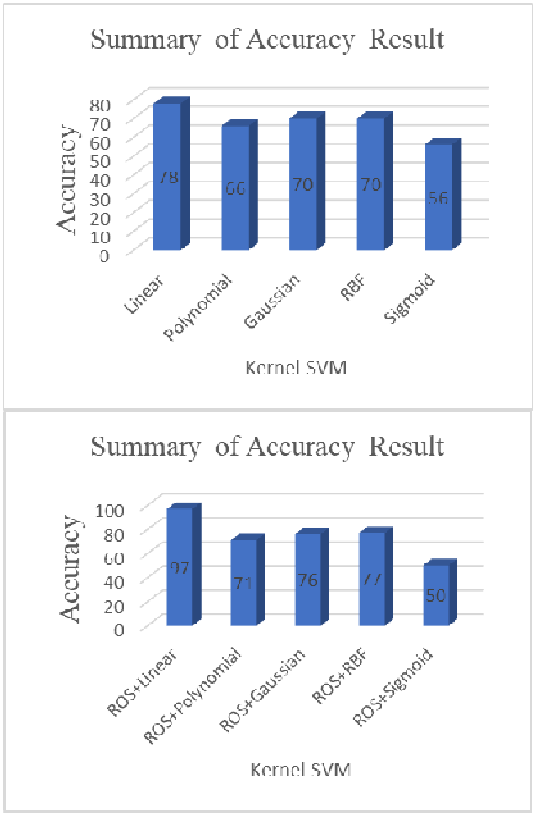

Fig. 6 Summary of Accuracy Result Schema for SVM

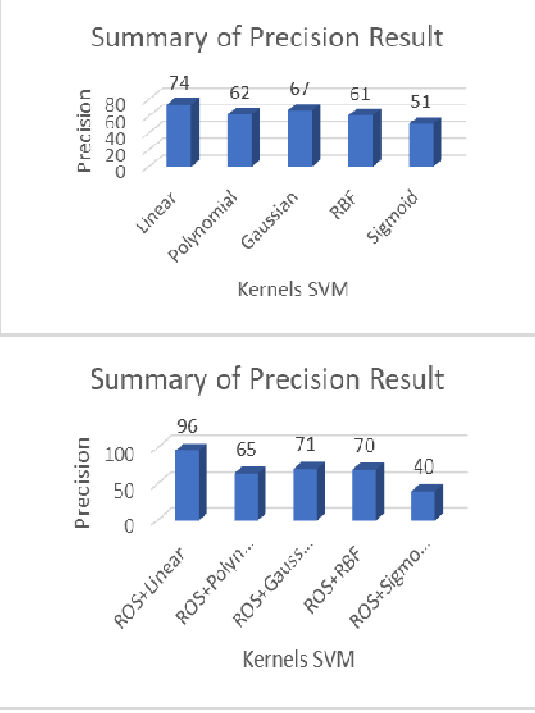

Fig. 7 Summary of Precision Result Schema for SVM

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, where the accuracy decreases by 6% when the oversampling technique is applied. This is because 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.

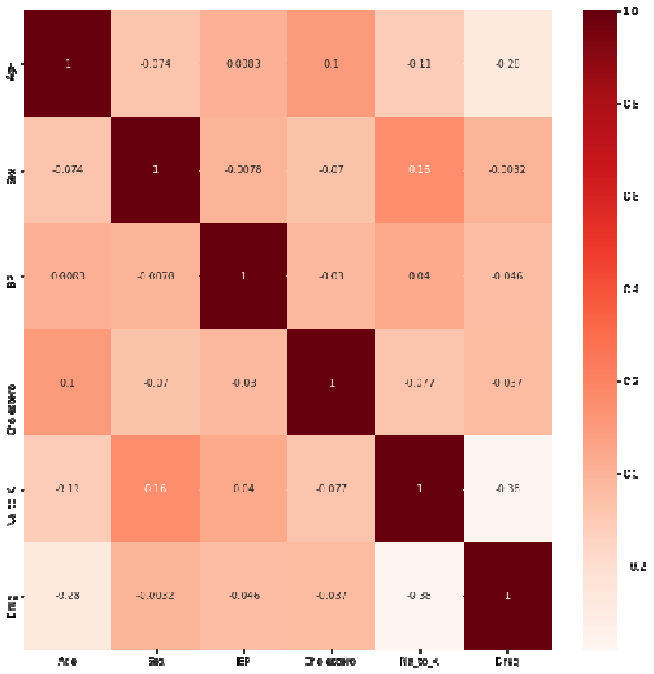

Fig. 8 Correlation Matrix Heatmap

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

[1] Li, Zhe, Yongtai Zhang, and Nianping Feng. "Mesoporous silica nanoparticles: Synthesis, classification, drug loading, pharmacokinetics, biocompatibility, and application in drug delivery." Expert opinion on drug delivery, vol. 16, no. 3, pp. 219-237, 2019, doi: 10.1080/17425247.2019.1575806.

[2] Nascimento, André CA, Ricardo BC Prudêncio, and Ivan G. Costa. "A multiple kernel learning algorithm for drug-target interaction prediction." BMC bioinformatics, vol. 17, no. 1, pp. 1-16, 2016, doi: 10.1186/s12859-016-0890-3.

[3] Yasuo, Nobuaki, Yusuke Nakashima, and Masakazu Sekijima. "Code-dti: Collaborative deep learning-based drug-target interaction prediction." 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018, pp. 792-797, doi: 10.1109/BIBM.2018.8621368.

[4] Olayan, Rawan S., Haitham Ashoor, and Vladimir B. Bajic. "DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches." Bioinformatics, vol. 34, no. 7, pp. 1164-1173, 2018, doi: 10.1093/bioinformatics/btx731.

[5] Ryu, Jae Yong, Hyun Uk Kim, and Sang Yup Lee. "Deep learning improves prediction of drug–drug and drug–food interactions." Proceedings of the National Academy of Sciences, vol. 115, no. 18, pp. E4304-E4311, 2018, doi: 10.1073/pnas.180329411.

[6] Shtar, Guy, Lior Rokach, and Bracha Shapira. "Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures." PloS one, vol. 14, no. 8, pp. e0219796, 2019, doi: 10.1371/journal.pone.0219796.

[7] Peska, Ladislav, Krisztian Buza, and Júlia Koller. "Drug-target interaction prediction: a Bayesian ranking approach." Computer methods and programs in biomedicine, vol. 15, no. 2, pp. 15-21, 2017, doi: 10.1016/j.cmpb.2017.09.003.

[8] Ezzat, Ali, et al. "Drug-target interaction prediction via class imbalance-aware ensemble learning." BMC bioinformatics, vol. 17, no. 19, pp. 267-276, 2016, doi: 10.1186/s12859-016-1377-y.

[9] Bagherian, Maryam, et al. "Machine learning approaches and databases for prediction of drug–target interaction: a survey paper." Briefings in bioinformatics, vol. 22, no. 1, pp. 247-269, 2021, doi: 10.1093/bib/bbz157.

[10] Chen, Ruolan, et al. "Machine learning for drug-target interaction prediction." Molecules, vol. 23, no. 9, pp. 2208, 2018, doi: 10.3390/molecules23092208.

[11] Lavecchia, Antonio. "Machine-learning approaches in drug discovery: methods and applications." Drug discovery today, vol. 20, no. 3, pp. 318-331, 2015, doi: 10.1016/j.drudis.2014.10.012.

[12] Liu, Yong, et al. "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction." PLoS computational biology, vol. 12, no. 2, pp. e1004760, 2016, doi: 10.1371/journal.pcbi.1004760.

[13] Peón, Antonio, Stefan Naulaerts, and Pedro J. Ballester. "Predicting the reliability of drug-target interaction predictions with maximum coverage of target space." Scientific reports 7.1, 1-11, Jun. 2017, doi: 10.1038/s41598-017-04264-w.

[14] He, Tong, et al. "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines." Journal of cheminformatics, vol. 9, no. 1, pp. 1-14, 2017, doi: 10.1186/s13321-017-0209-z.

[15] Mohammed Nazim Uddin, Md. Ferdous Bin Hafiz, Sohrab Hossain and Shah Mohammad Mominul Islam, "Drug Sentiment Analysis using Machine Learning Classifiers" International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130112.

[16] Wongyikul, Pakpoom, et al. "High alert drugs screening using gradient boosting classifier." Scientific Reports, vol. 11, no. 1, pp. 1-24, Oct. 2021, doi: 10.1038/s41598-021-99505-4.

[17] Shanbhag, Shrinivas V., et al. "Drug-Drug Interaction Extraction Based on Deep Learning Models." Soft Computing for Problem Solving. Springer, Singapore, pp. 691-706, 2021, doi: 10.1007/978-981-16-2709-5_53.

[18] Liu, Bin, et al. "Drug-target interaction prediction via an ensemble of weighted nearest neighbors with interaction recovery." Applied Intelligence 52.4, 3705-3727, 2022, doi: 10.1007/s10489-021-02495-z.

[19] Thafar, Maha A., et al. "DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques." Journal of Cheminformatics 12.1, 1-17. 2020. doi: 10.1186/s13321-020-00447-2.

[20] Pliakos, Konstantinos, and Celine Vens. "Drug-target interaction prediction with tree-ensemble learning and output space reconstruction." BMC bioinformatics, vol. 21, no. 1, pp. 1-11. 2020, doi: 10.1186/s12859-020-3379-z.

[21] Mohan, Maya. "Ensemble Learning Models for Drug Target Interaction Prediction." 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 16 June, 2022, doi: 10.1109/ICAAIC53929.2022.9793081.

[22] Ye, Qing, Xiaolong Zhang, and Xiaoli Lin. "Drug-target interaction prediction via multiple output deep learning." 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, January, 2020, doi: 10.1109/BIBM49941.2020.9313488.

[23] Mahmud, SM Hasan, et al. "Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting." Analytical biochemistry 589, pp. 113507, 2020, doi: 10.1016/j.ab.2019.113507.

[24] Mahmud, SM Hasan, et al. "iDTi-CSsmoteB: identification of drug–target interaction based on drug chemical structure and protein sequence using XGBoost with over-sampling technique SMOTE." IEEE Access 7, pp. 48699-48714, 2019, doi: 10.1109/ACCESS.2019.2910277.

[25] Branco, Paula, Luís Torgo, and Rita P. Ribeiro. "A survey of predictive modeling on imbalanced domains." ACM Computing Surveys (CSUR) 49.2, pp. 1-50, 2016, doi: 10.1145/2907070.

[26] Nur Ghaniaviyanto Ramadhan, Adiwijaya and Ade Romadhony, "Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest" International Journal of Advanced Computer Science and Applications(IJACSA), 12(7), 2021, doi: 10.14569/IJACSA.2021.0120726.

[27] Ertekin, Şeyda. "Adaptive oversampling for imbalanced data classification." Information Sciences and Systems 2013. Springer, Cham 264, pp. 261-269, 2013, doi: 10.1007/978-3-319-01604-7_26.

[28] Zhang, Yongli. "Support vector machine classification algorithm and its application." International conference on information computing and applications (ICICA). Springer, Berlin, Heidelberg, 308, May, 2012, doi: 10.1007/978-3-642-34041-3_27.

[29] Ramadhan, Nur Ghaniaviyanto, and Azka Khoirunnisa. "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine." Jurnal Media Informatika Budidarma 5.4, pp. 1580-1584, 2021, doi: 10.30865/mib.v5i4.3347.

[30] Ramadhan, Nur Ghaniaviyanto, and Teguh Ikhlas Ramadhan. "Analysis Sentiment based on IMDB aspects from movie reviews using SVM." Sinkron: jurnal dan penelitian teknik informatika 7.1, 39-45, 2022, doi: 10.33395/sinkron.v7i1.11204.