BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0141–0150
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0141-0150

APPLICATION OF SUPPORT VECTOR MACHINE FOR CLASS IMBALANCE LEARNING TO PREDICT ANTICANCER COMPOUNDS OF MEDICINAL PLANTS IN WEST SULAWESI

Hikmah[1], Nur Hilal A. Syahrir[2*], Putri Indi Rahayu[3]

[1,2,3] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat
St. Prof. Dr. Baharuddin Lopa, Majene, 91412, Indonesia.

*Corresponding author's e-mail: * **nurhilal.asyahrir@unsulbar.ac.id***

ABSTRACT

Article History:

Received: 31th July 2023
Revised: date, month year
Accepted: date, month year

Keywords:

ADASYN;
DB-SMOTE;
Imbalance;
SMOTE;
Oversampling;
Undersampling.

Indonesian medicinal plants, such as turmeric and soursop, have shown promising anticancer properties through their bioactive compounds, like curcumin and extracts from soursop. Despite many extensive studies on medicinal plants in Indonesia, research revealing the activity of natural products in West Sulawesi is still limited, and the studies focus mainly on ethnobotanical research. In this work, we propose a machine-learning approach to predict the anticancer activity of compounds in medicinal plants in West Sulawesi by leveraging high throughput-screening data, especially molecular information from a public database. We applied a Support Vector Machine (SVM) with five sampling techniques to address data imbalance. We also evaluated the performance in selecting the best combination in handling class imbalance learning in our dataset. The result shows that undersampling and ADSYN methods can improve the prediction of anticancer activity. Based on the two methods of balancing data, we have ten potential anticancer compounds from three medicinal plants in West Sulawesi.

How to cite this article:

Hikmah, N. H. A. Syahrir and P. I. Rahayu., "APPLICATION OF SUPPORT VECTOR MACHINE FOR CLASS IMBALANCE LEARNING TO PREDICT ANTICANCER COMPOUNDS OF MEDICINAL PLANTS IN WEST SULAWESI," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0141-0150, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Cancer is one of the primary causes of death, with millions of people diagnosed annually in developed and developing countries, including Indonesia. In 2018, around 396,914 new cancer cases and almost 234,511 cancer-related deaths were estimated to happen in Indonesia, according to the Global Cancer Observatory (GLOBOCAN) [1]. Despite advancements in conventional cancer therapies, the effectiveness of conventional drugs is often limited by toxic side effects, which often impact patients' quality of life. In contrast, medicinal plants offer a potential avenue to discover naturally occurring compounds with more advantageous safety profiles. Medicinal plants have garnered considerable attention in cancer research due to their potential as a source of novel and effective anticancer agents. These plants contain many bioactive compounds, such as alkaloids, flavonoids, terpenoids, and polyphenols, demonstrating anticancer properties in preclinical studies.

The discovery of anticancer compounds in various medicinal plants has been successfully done. For instance, turmeric (*Curcuma longa*) has been extensively studied for its anti-inflammatory and anticancer effects, which are attributed to its main bioactive compound, curcumin [2]. Another example is the extract from soursop (*Annona muricata*), which has shown selective cytotoxicity against cancer cells, indicating its potential as an alternative or adjunct to conventional cancer treatments [3]. As a country with rich medicinal plant diversity, Indonesia presents a promising avenue for discovering new and effective anticancer agents by combining traditional knowledge with modern scientific approaches.

Although many studies on medicinal plants and their natural products in Indonesia have been conducted, research on natural products of medicinal plants in West Sulawesi is still limited. The research of medicinal plants in West Sulawesi focused on ethnobotanical research, which only identified and listed the family, species, local name, plant habitus, parts of the plant used, preparation of medicinal plants, and types of diseases/disorders or efficacy [4]–[6]. The potential compound in medicinal plants in West Sulawesi has yet to be revealed. Therefore, in this research, we aim to predict compounds' anticancer activity in West Sulawesi's medicinal plants using a machine learning model to reveal the potential compounds in medicinal plants.

We proposed a machine-learning approach using information on compounds in West Sulawesi medicinal plants. We utilized high throughput screening from public databases to collect information on the compounds, gathered the molecular details on the compound, and constructed features for the machine learning model. Potential anticancer compounds in medicinal plants in West Sulawesi can be efficiently identified using a machine-learning model for medicinal plants. Machine learning (ML) algorithms have successfully been used for classification tasks to predict compound activity. This work has the advantage of the Support Vector Machine (SVM) due to its effectiveness in handling both low-dimensional and high-dimensional data without the need to assume a specific relationship between bioactivity and molecular features [7]. However, due to the imbalance problem in the dataset, we combined SVM with several techniques for handling imbalance to predict anticancer activity in medicinal plants.

2. RESEARCH METHODS

2.1 Data Sources

This research used data from medicinal plants, which are well-known as herbal or alternative medicine. From medicinal plants, we collected compound data especially chemical structure data in two dimensions. Based on the literature review, we selected three medicinal plants: *Crassocephalum crepidioides*, *Elephantopus scaber*, and *Imperata cylindrica*. Those three medicinal plants are usually used as traditional medicine in treating cancer in West Sulawesi [8]. To build a classification model, we also collected drug compounds from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) [9]. All the training compounds we collected have Anatomical Therapeutic Chemical (ATC) codes in the PubChem database. The Anatomical Therapeutic Chemical (ATC) classification system generates ATC codes, where the compounds or drugs are divided into groups based on the organ or system on which the compounds act. In order to predict whether the *Crassocephalum crepidioides*, *Elephantopus scaber*, and *Imperata cylindrical contain* antibacterial compounds, we searched the compounds of those three medicinal plants in the PubChem database before using them as our testing data. In the PubChem database, we used the name of plants on the web and select the “Taxonomy” in order to find all information about the plants including natural products contained in the plants which have been identified by previous research.

2.2 Methods

2.2.1 Labelling and Feature Construction in Training Data

All compounds with ATC codes are classified into two classes, i.e., negative and positive classes. Compounds in the positive class are compounds with anticancer activity and vice versa. Since the number of compounds related to anticancer is relatively small (minority class) compared with non-anticancer compounds (majority class), we handled this imbalance problem in our analysis. We generated chemical similarity of all compounds to construct our learning feature or attribute. We calculated the chemical similarity of compounds using Tanimoto coefficients [10] based on binary data in fingerprint data of compounds.

2.2.2 Pre-processing Data

We applied five independent techniques in handling our imbalance learning problems in this stage.

- Undersampling. This method works by eliminating some data from the majority class, which is done randomly. Data from the majority class are reduced to be balanced with the minority class. Techniques for reducing data in undersampling are very varied. Several methods focus on the majority class, but some focus on both classes only in the border area [11], [12].
- Oversampling. The technique works by replicating data selected from a small class called a minority. The oversampling results cannot constantly improve the predictions of the minor class. If minor class data is replicated in large quantities, it will be difficult to identify data with similar characteristics but in different classes [11], [13].
- Synthetic Minority Oversampling Technique (SMOTE). The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method used to increase the number of minority (positive) classes in a dataset. It involves replicating random data from the minority class to match the quantity of the majority class. The main idea behind this approach is to create replicas of the minority data. To implement the SMOTE algorithm, the k-nearest neighbour (KNN) for each minority class is identified, and synthetic data is then generated by duplicating the minority class instances in proportion to a desired percentage and randomly selecting the KNN. While SMOTE can enhance accuracy in minority classes, it may lead to overgeneralization, as the newly generated data could spread across both minority and majority classes [14], [15].
- Density-Based Synthetic Minority Oversampling Technique (DB SMOTE). DB SMOTE utilizes the principles of density-based spatial clustering of applications with noise (DBSCAN). It generates synthetic instances by establishing the shortest path between each positive instance and the pseudo-centroid cluster within the minority class. Consequently, the synthetic dataset becomes dense around the centroid of the original positive cases group [16], [17].
- Adaptive Synthetic Sampling Approach (ADASYN). The ADASYN approach addresses the issue of imbalanced data by employing distribution weights to oversample minority classes. This method aims to balance the dataset by increasing the representation of the minority classes, thus mitigating the problem of data imbalance. ADASYN has parameters that are used to determine the expected equilibrium level of (β), and a limit set as the maximum tolerance degree of the imbalance ratio of the class (d_{th}) [18].

2.2.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification. It aims to find an optimal hyperplane that maximizes the margin between two classes in a high-dimensional feature space, making it effective for both linear and non-linear data classification. By using support vectors and applying the kernel trick, SVM can handle non-linearly separable data by transforming it into a higher-dimensional space. SVM's advantages include its ability to handle high-dimensional data, model complex decision boundaries, and its resilience to overfitting with proper parameter tuning [19]. Linearly separable data is data that can be separated linearly. For example, $x_i = \{x_i \cdots, x_n\}$, $x_i \in \Re^n$ is a dataset and$y_i \in \{+1, -1\}$ is the label of the dataset x_i. SVM works to find the best separator factor that can separate data and has the largest margin.

Data points that are in the bounding field are called support vectors. The bounding field that separates two different classes so that the two classes can be written the form of the equation as follows:

$$\begin{aligned} \boldsymbol{x_i w} + b \geq +1, y_i = +1 \\ \boldsymbol{x_i w} + b \leq -1, y_i = -1 \end{aligned} \tag{1}$$

The variable $\boldsymbol{w}$ is a normal field, and b is biased. Margin calculation with bounding fields as follows:

$$\frac{1 - b - (1 - b)}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|} \tag{2}$$

If the maximum is from $\frac{2}{\|\boldsymbol{w}\|}$ an optimal hyperplane is tantamount to minimizing $\frac{1}{2}\|\boldsymbol{w}\|^2$. The limiting field of **Equation (1)** can be formed into the following inequality:

$$y_i(\boldsymbol{x_i w} + b) - 1 \geq 0 \tag{3}$$

So that the formulation of optimization problems in SVM is as follows:

$$\begin{gathered} min \frac{1}{2}\|w\|^2 \\ \text{with } y_i(\boldsymbol{x_i w} + b) - 1 \geq 0 \end{gathered} \tag{4}$$

Unlike if the data owned cannot be separated linearly so that a solution to the problem is needed, namely by using the kernel or "Kernel Trick". This method transforms data into a higher dimensional form or feature space that can later be separated linearly in the feature space. Some kernel functions that can be used are presented in **Table 1** below **[20]**:

Table 1. Kernel Function

Kernel	Function
Linear	$K(x_i, x) = {x_i}^t x$
Polynomial	$K(x_i, x) = (\gamma {x_i}^t x + r)^p, \gamma > 0$
Radial Basis Function (RBF)	$K(x_i, x) = exp(-\gamma\|x - x_i\|^2)$
Sigmoid	$K(x_i, x) = tanh(\gamma {x_i}^t x + r)$

This SVM and kernel method is then used to modelling drug compounds by using matrix of chemical similarity as the input and category variable explaining the compounds is anticancer or not as the target.

2.2.3 Performance Evaluation

Assessment criteria are the most essential aspect in evaluating classification performance. The simplest way to assess classification performance is to cross-tabulate the actual and predicted classes. The result of this cross-tabulation is known as the confusion matrix (**Table 2**). Regarding two classes, the confusion matrix provides information regarding the number of predicted classes in the column and the actual number of courses in the row.

Table 2. Confusion Matrix

Class		**Prediction**	
		Majority Class	**Minority Class**
Actual	**Majority Class**	TP (True Positive)	FN (False Negative)
	Minority Class	FP (False Positive)	TN (True Negative)

Performance evaluation can be calculated based on accuracy, specificity, sensitivity, balanced accuracy, G-mean, F1 Score, and Area Under Curve (AUC) **[21]** with the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{8}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

$$FPR = 1 - Specificity \tag{13}$$

$$AUC = \frac{1 + Sensitivity - FPR}{2} \tag{14}$$

Accuracy is used to measure how accurate the model is in making a prediction. Specificity is used to measure how accurately a model predicts minority classes. Sensitivity is used to measure or proportion accuracy in predicting the majority class. Balanced accuracy is the average of the sensitivity and specificity values. G-mean is the geometric mean of sensitivity and specificity and is worth one if all observations can be classified precisely. The F1 score is the harmonic average of the recall and precision values. AUC is the single measure used in evaluating the best model.

2.2.4 Predicting Anticancer Activity in Medicinal Plants

We select the best classification model in the previous stage to predict anticancer compounds in medicinal plants. The feature of compounds in medicinal plants is constructed with the same method in the training set.

3. RESULTS AND DISCUSSION

3.1 Labelling and Feature Construction of Compounds Data for Training Data

There are 4560 compounds collected with Anatomical Therapeutic Chemical (ATC) codes in the PubChem database. 30 compounds from 4530 compounds are not valid. We excluded the compounds in further analysis. We used the category of “antineoplastic and immunomodulating agents” in ATC codes, which are L01 codes, as a positive class since compounds are related to anticancer activity. Antineoplastic agents, also known as anticancer or chemotherapy drugs, are medications used to treat various types of cancer. They target and destroy cancer cells or inhibit their growth and division [22]. Immunomodulating agents, on the other hand, are a distinct group of drugs that work by modifying the body's immune response. Immunomodulating agents can potentially augment the immune system's capacity to identify and combat cancer cells [23]. These agents are commonly utilized in immunotherapy, an advanced strategy for treating cancer that revolves around strengthening the body's innate ability to defend itself against cancer. Therefore, 424 compounds with antineoplastic and immunomodulating agents were labeled as a positive class, and the rest were labeled as a negative class. We gained a symmetrical chemical similarity matrix with 4530 rows and 4530 columns using the Tanimoto Coefficient. In other words, we have an equal number of instances and attributes or features. The range of values is zero to one, so we did not use transformation or normalization in the dataset.

3.2 Pre-processing Data

Our original data contained 4106 negative compounds and 424 positive compounds. As shown in Table 3, after splitting training and testing, we have applied balanced techniques in only the training set to avoid synthetic data being used as a testing set.

Table 3. The Number of Instances Using Balanced Techniques in Data Pre-processing

Data Splitting	Balanced Techniques	Class		Ratio
		Negative	Positive	
Training Set (80%)	Original	3285	340	1:9.66
	Undersampling	340	340	1:1
	Over-sampling	3285	3285	1:1
	SMOTE	3285	3060	1:107
	DB-SMOTE	3285	3076	1: 1.08
	ADASYN	3285	3258	1: 1.01
Testing Set (20%)	-	821	84	1:9.77

Overall, all the techniques in balancing data reduce the ratio of the two classes from 1:9.66 to nearly 1:1. In addition, it can be seen that the ratio of our testing set is approximately close to the characteristic of the original training set and the whole dataset.

3.3 Modelling and Performance Evaluation of Support Vector Machine

We applied SVM with four types of kernel functions, i.e., linear, polynomial, radial basis function, and sigmoid, to find the optimal hyperplane in our original dataset. The result of the testing set in the SVM model with kernels is shown in **Table 4**. The result indicates that the sigmoid function has the lowest number in all our evaluation criteria. Mapping this dataset into a higher-dimensional space using sigmoid functions is not appropriate. The other kernel function that maps data into high dimensions is polynomial, which has the highest accuracy and perfect sensitivity. However, the SVM model of this function cannot effectively distinguish the minority class. As our primary purpose in this research is to predict the minority class, we excluded the polynomial function, which is ineffective in predicting the minority class.

Table 4. Evaluation of Testing Set Using Model from Original Training Set in Four Types of Kernel Function

Metric	Kernels			
	Linear	Polynomial	RBF	Sigmoid
Accuracy	0.92	0.94	0.94	0.87
Balanced Accuracy	**0.78**	0.67	0.66	0.51
Sensitivity	0.95	1.00	1.00	0.95
Specificity	**0.61**	0.33	0.32	0.07
Precision	0.96	0.94	0.94	0.91
F1	0.96	0.97	0.97	0.93
AUC	**0.78**	0.67	0.66	0.51
G-mean	**0.76**	0.58	0.57	0.26

The same result of the polynomial function is the radial function which also has low sensitivity. Although the radial kernel is the most widely used in SVM due to its suitability to complex and nonlinear patterns, the radial kernel has poor performance in this study. Data points in our dataset tend to show a linear pattern in the original feature space. Therefore, we used linear kernel further analysis because it works more efficiently and provides the best results (balanced accuracy, specificity, AUC, and G-mean) compared to the other kernel function. One advantage of using linear kernels is that they are computationally less expensive than other kernel functions.

The result of several sampling techniques in balancing data is shown in **Table 5**. Compared with the original dataset, the method of balancing data can improve all evaluation metrics, although the improvement of all metrics is relatively slight except the undersampling method. The specificity of the undersampling method is significantly higher than the original data, but unfortunately, the sensitivity of this method is worse than other balancing techniques. In other words, although the capability to predict minority class improves, at the same time, SVM with undersampling becomes less effective in excluding minority classes in positive class, affecting its accuracy.

Table 5. **Evaluation of Testing Set of Linear SVM Using Balanced Techniques**

Dataset	Evaluation Metric				
	Accuracy	**Balanced Accuracy**	**Sensitivity/Recall**	**Specificity**	**Precision**
Original	0.92	0.78	0.95	0.61	0.96
Undersampling	0.74	**0.80**	0.73	**0.87**	**0.98**
Oversampling	0.92	**0.80**	0.95	0.64	0.96
SMOTE	0.92	**0.80**	0.95	0.64	0.96
DB-SMOTE	0.92	0.79	0.95	0.64	0.96
ADASYN	**0.93**	0.79	**0.96**	0.62	0.96

High specificity becomes essential in this study because the study aims to predict the minority class. Therefore, the undersampling method can be considered a good model for predicting anticancer compounds in medicinal plants.

In this imbalanced classification problem, where negative anticancer compounds are significantly more prevalent than anticancer compounds, standard accuracy metrics may not accurately represent our SVM's model performance. We used other evaluation metrics to evaluate our model comprehensively i.e.: the F1 score, AUC-ROC (Area Under the Receiver Operating Characteristic curve), and G-mean (Geometric Mean) commonly used in imbalanced classification tasks. The result of all metrics is presented in **Table 6**. According to **Table 6**, the highest G-mean and AUC scores are gained when using SVM with undersampling. As previously discussed, the undersampling method has limitations in predicting minority class, and we can see that the other method that can be considered in the prediction task is SVM with the ADASYN method, which can effectively distinguish minority class better than the undersampling method. ADASYN has the best sensitivity and F1 score. The ADASYN method effectively excluded positive anticancer activities in the negative category. Thus, we used both resampling techniques in predicting new compounds in this work.

Table 6. **Imbalance Evaluation Criteria of Testing Set of Linear SVM Using Balanced Techniques**

Dataset	Evaluation Metric		
	F1	**G-mean**	**AUC**
Original	0.9554	0.7600	0.7792
Undersampling	0.8347	**0.7943**	**0.7975**
Oversampling	0.9571	0.7820	0.7971
SMOTE	0.9558	0.7810	0.7959
DB-SMOTE	0.9545	0.7800	0.7946
ADASYN	**0.9610**	0.7713	0.7900

3.4 Predicting Anticancer Activity in Medicinal Plants

There are 266 compounds that we collected through intensive searching in the PubChem database from three medicinal plants: *Crassocephalum crepidioides, Elephantopus scaber*, and *Imperata cylindrica*. The number of compounds in these medicinal plants could increase if additional compounds are identified in those plants. The activity of the majority of compounds from 266 compounds is still unknown. Therefore, we aim to reveal what compounds play a role in treating cancer considering those three medicinal compounds have been used for a long time as traditional medicine for anticancers in Sulawesi West. We tested 266 compounds in our anticancer classification modelling in order to obtain the potential compound to be the anticancer compounds.

Table 7. Class Prediction of Medicinal Plants' Compound in West Sulawesi

Medicinal Plants	Number of Compounds	Class Prediction			
		Undersampling SVM		ADASYN SVM	
		Negative	Positive	Negative	Positive
Crassocephalum crepidioides	86	34	52	80	6
Elephantopus scaber	31	20	11	30	1
Imperata cylindrica	149	115	34	144	5
Total	266	169	97	254	12

We used SVM with undersampling and ADASYN methods as the best model to predict anticancer activity in medicinal plants. We compare the result of anticancer compounds classification which is shown in Table 7. The number of compounds predicted as anticancer compounds (positive class) and non-anticancer compounds (negative class) is extremely different. According to Table 7, the number of predicted anticancer compounds using SVM with ADASYN is significantly smaller than using SVM with undersampling. The ADASYN method effectively identifies non-anticancer activity (negative class). In contrast, the undersampling method effectively identifies anticancer compounds (positive class).

From 97 compounds predicted as positive compounds using undersampling SVM, 87 compounds were predicted as negative compounds using ADASYN SVM. Thus, ten compounds were predicted as positive compounds in both methods. The ten compounds are the potential anticancer compounds we obtain in this work. Table 8 shows ten potential anticancer compounds based on the SVM model—most compounds from *Crassocephalum crepidioides* and only one compound from *Elephantopus scaber*. *Imperata cylindrica* only has three potential compounds, although the plant has the highest number of identified compounds.

The potential compounds are still unexplored as anticancer compound. Therefore, this result can be used as a very early drug discovery step in searching anticancer compounds to be validated in the wet laboratory.

Table 8. Ten Potential Anticancer Compounds from Medicinal Plants in West Sulawesi

CID	Sources of Medicinal plants	Compound Name
10319743	*Crassocephalum crepidioides*	5-epi-Vibsanin C
10432070	*Crassocephalum crepidioides*	Vibsanin C
93488714	*Crassocephalum crepidioides*	[(E)-2-[(1S,2S,7S)-2-[(3S)-3,4-dihydroxy-4-methylpentyl]-5-(hydroxymethyl)-2-methyl-6-oxo-7-(2-oxopropyl)cyclohept-4-en-1-yl]ethenyl] 3-methylbut-2-enoate
93492969	*Crassocephalum crepidioides*	[(E)-2-[(1S,2S,7S)-5-(hydroxymethyl)-2-[(3R)-3-hydroxy-4-methylpent-4-enyl]-2-methyl-6-oxo-7-(2-oxopropyl)cyclohept-4-en-1-yl]ethenyl] 3-methylbut-2-enoate
162966824	*Crassocephalum crepidioides*	[(E)-2-[(1R,2R,7R)-2-[(3S)-3,4-dihydroxy-4-methylpentyl]-5-(hydroxymethyl)-2-methyl-6-oxo-7-(2-oxopropyl)cyclohept-4-en-1-yl]ethenyl] 3-methylbut-2-enoate
163005160	*Crassocephalum crepidioides*	2-[(1S,2R,3S,9S)-3,9-dimethyl-3-(4-methylpent-3-enyl)-8,12-dioxatricyclo[7.2.1.01,6]dodec-5-en-2-yl]ethenyl 3-methylbut-2-enoate
162852041	*Elephantopus scaber*	(1S,5S,6R,8S,11R)-8-hydroxy-1,5,11-trimethyl-9-oxotricyclo[6.2.1.02,6]undec-2-ene-11-carbaldehyde
5280450	*Imperata cylindrica*	Linoleic Acid
5282457	*Imperata cylindrica*	Linoelaidic acid
101915994	*Imperata cylindrica*	Graminonea

4. CONCLUSIONS

1. Support Vector Machine in chemical similarity data is more suitable for using linear functions rather than non-linear functions to predict anticancer compounds. Other machine-learning models are still required to obtain better performance in this dataset.
2. Balancing techniques such as undersampling and ADASYN before applying the SVM method successfully increase the performance in predicting anticancer compounds in imbalance class data. However, there is a limitation in both techniques due to the low sensitivity or specificity. Many sampling techniques, especially improvement of SMOTE techniques, can be conducted in future research to increase sensitivity or specificity simultaneously.
3. There are ten potential anticancer compounds in West Sulawesi from three medicinal plants (*Crassocephalum crepidioides, Elephantopus scaber,* and *Imperata cylindrica)* obtained using SVM with balancing techniques. These potential compounds can be used and analyzed in future research to ensure the biological activity of compounds rather than utilizing another compound with a low possibility of having anticancer properties. Using potential compounds through computational methods before validating them in the wet laboratory can efficiently reduce the cost, time, and human resources.

ACKNOWLEDGMENT

This research is financially supported by LPPM Universitas West Sulawesi through the "Penelitian Dasar" scheme.

REFERENCES

[1] B. Andinata, A. Bachtiar, P. Oktamianti, J. R. Partahi, and M. S. A. Dini, "A Comparison of Cancer Incidences Between Dharmais Cancer Hospital and GLOBOCAN 2020: A Descriptive Study of Top 10 Cancer Incidences," *Indonesian Journal of Cancer*, vol. 17, no. 2, pp. 119–122, 2023.

[2] A. Zia, T. Farkhondeh, A. M. Pourbagher-Shahri, and S. Samarghandian, "The role of curcumin in aging and senescence: Molecular mechanisms," *Biomedicine & Pharmacotherapy*, vol. 134, p. 111119, 2021.

[3] S. Ilango *et al.*, "A review on annona muricata and its anticancer activity," *Cancers (Basel)*, vol. 14, no. 18, p. 4539, 2022.

[4] G. M. Nurdin, A. P. Sari, and H. Herni, "Identifikasi Tumbuhan Obat Masyarakat Desa Pao-Pao Kabupaten Polewali Mandar Provinsi Sulawesi Barat," *Biosfer: Jurnal Biologi dan Pendidikan Biologi*, vol. 7, no. 1, pp. 20–29, 2022.

[5] H. Hastuti, I. Lestari, M. Yunus, and A. Hasyim, "Inventarisasi Tumbuhan Berkhasiat Obat di Desa Pokkang, Kec. Kalukku, Kabupaten Mamuju, Provinsi Sulawesi Barat," *Jurnal Biosense*, vol. 5, no. 01, pp. 41–54, 2022.

[6] H. Alang, S. Rosalia, and A. D. R. Ainulia, "Inventarisasi tumbuhan obat sebagai upaya swamedikasi oleh masyarakat suku mamasa di Sulawesi Barat," *Quagga: Jurnal Pendidikan Dan Biologi*, vol. 14, no. 1, pp. 77–87, 2022.

[7] R. Zhang, X. Li, X. Zhang, H. Qin, and W. Xiao, "Machine learning approaches for elucidating the biological effects of natural products," *Nat Prod Rep*, vol. 38, no. 2, pp. 346–361, 2021.

[8] S. Syamsiah, H. Karim, A. F. Arsal, and S. Sondok, "Kajian Etnobotani dalam Pemanfaatan Tumbuhan Obat Tradisional di Kecamatan Pana Kabupaten Mamasa, Sulawesi Barat," *Jurnal Bionature*, vol. 22, no. 2, pp. 1–12, 2021.

[9] S. Kim *et al.*, "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Res*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[10] A. Rácz, D. Bajusz, and K. Héberger, "Life beyond the Tanimoto coefficient: Similarity measures for interaction fingerprints," *J Cheminform*, vol. 10, no. 1, pp. 1–12, 2018, doi: 10.1186/s13321-018-0302-y.

[11] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th international conference on information and communication systems (ICICS)*, IEEE, 2020, pp. 243–248.

[12] S. Liu and K. Zhang, "Under-sampling and feature selection algorithms for S2SMLP," *IEEE Access*, vol. 8, pp. 191803–191814, 2020.

[13] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, "Classification of imbalanced data by oversampling in kernel space of support vector machines," *IEEE Trans Neural Netw Learn Syst*, vol. 29, no. 9, pp. 4065–4076, 2017.

[14] A. Ishaq *et al.*, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE access*, vol. 9, pp. 39707–39716, 2021.

[15] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach Learn*, pp. 1–21, 2023.

[16] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: density-based synthetic minority over-sampling technique," *Applied Intelligence*, vol. 36, pp. 664–684, 2012.

[17] C.-K. Ma and Y.-J. Park, "A new instance density-based synthetic minority oversampling method for imbalanced classification problems," *Engineering Optimization*, vol. 54, no. 10, pp. 1743–1757, 2022.

[18] J. Brandt and E. Lanzén, “A comparative review of SMOTE and ADASYN in imbalanced data classification,” 2021.

[19] D. A. Pisner and D. M. Schnyer, “Support vector machine,” in *Machine learning*, Elsevier, 2020, pp. 101–121.

[20] A. Patle and D. S. Chouhan, “SVM kernel functions for classification,” in *2013 International conference on advances in technology and engineering (ICATE)*, IEEE, 2013, pp. 1–9.

[21] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, “Cross-validation metrics for evaluating classification performance on imbalanced data,” in *2019 international conference on computer, control, informatics and its applications (IC3INA)*, IEEE, 2019, pp. 14–18.

[22] M. J. Nime *et al.*, “Studies on Antioxidant and Antineoplastic Potentials of Oldenlandia corymbosa Linn. Leaves,” *Journal of Fundamental and Applied Pharmaceutical Science*, vol. 3, no. 2, p. 84, 2023.

[23] M. Zebeaman, M. G. Tadesse, R. K. Bachheti, A. Bachheti, R. Gebeyhu, and K. K. Chaubey, “Plants and Plant-Derived Molecules as Natural Immunomodulators,” *Biomed Res Int*, vol. 2023, 2023.