Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 22, No. 1, November 2022, pp. 13∼24
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v22i1.1803

The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data

Muhammad Ibnu Choldun Rachmatullah
Politeknik Pos Indonesia, Bandung, Indonesia

Article Info

Article history:

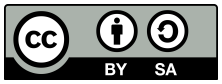
Received February 02, 2022
Revised July 07, 2022
Accepted October 10, 2022

Keywords:

Imbalanced data
Oversampling
Classification
Multi Class
Repeated SMOTE

ABSTRACT

One of the problems that are often faced by classifier algorithms is related to the problem of imbalanced data. One of the recommended improvement methods at the data level is to balance the number of data in different classes by enlarging the sample to the minority class (oversampling), one of which is called The Synthetic Minority Oversampling Technique (SMOTE). SMOTE is commonly used to balance data consisting of two classes. In this research, SMOTE was used to balance multi-class data. The purpose of this research is to balance multi-class data by applying SMOTE repeatedly. This iterative process needs to be applied if the number of unbalanced data classes is more than two classes, because the one-time SMOTE process is only suitable for binary classification or the number of unbalanced data classes is only one class. To see the performance of iterative SMOTE, the SMOTE datasets were classified using a neural network, k-NN, Nave Bayes, and Random Forest and the performance measures were measured in terms of accuracy, sensitivity, and specificity. The experiment in this research used the Glass Identification dataset which had six classes, and the SMOTE process was repeated five times. The best performance was achieved by the Random Forest classifier method with accuracy = 86.27%, sensitivity = 86.18%, and specificity = 95.82%. The result of experiment present that repeated SMOTE results can increase the performance of classification.



Corresponding Author:

Muhammad Ibnu Choldun Rachmatullah, 081347595733,
D III Information System,
Politeknik Pos Indonesia, Bandung, Indonesia,
Email: ibnucholdun@poltekpos.ac.id

How to Cite: M. Rachmatullah, The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data, MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 1, pp. 13-24, Nov. 2022.

1. INTRODUCTION

The problem of classification of imbalanced data occurs in various fields and has given a lot of attention from researchers. For example, imbalanced data related to anomaly detection [1], fault diagnosis kesalahan [2, 3], disease diagnosis [4, 5], or facial recognition [6]. Imbalanced data [7] means that in the classification process involving two classes, the number of samples for a particular class is much larger than the samples number for other classes [8, 9]. Usually, the class that has the larger number of samples is called the majority class, and the other class is called the minority. Several research have provided that when the number of certain classes is much larger than in other classes, the existing classification method will provide more focus to the accuracy of classification of the majority class sample in the classification process than the minority class. Several research have provide that when the sample size in the majority class is three or more times the size of sample in the minority class, can usually be said as imbalanced dataset. Whereas in some cases, minority class samples in imbalanced datasets are often the focus of research. So, a strategy is needed to increase the minority class samples accuracy [10].

Recently, the classification of imbalanced datasets has been the focus of attention by researchers. An imbalanced class distribution can obscure the results of machine learning or data mining algorithms, because the performance measure used by machine learning algorithms is usually overall accuracy. For example, there are 90 normal samples in the disease classification and only 10 abnormal samples, although all abnormal samples are misclassified, the accuracy of the model is still 90% [11, 12]. Imbalanced class can interfere with the ability of prediction of algorithms or classification methods because algorithms pursue classification accuracy as a whole [13]. To solve classification problems on imbalanced data, researchers can increase the performance of classification algorithms at the data level or algorithm level [13, 14]. One of the improvement technique at the level of data is to balance the number of data in different classes by adding samples to the minority class (oversampling) or removing samples from the majority class (undersampling) [15, 16]. The advantage of improving at the data level is that the method does not need to be limited by a specific domain and classifier model [17], and is more commonly applied than improving the algorithm to fit a specific classifier [15]. The undersampling method reduces the imbalanced dataset by decrease the sample of class with a larger sample size. This decreasing can be done randomly, and usually this is often called random undersampling [18]. While this method can balance the number of other sample classes and reduce the total sample size, which can reduce computational time, there is a risk of losing significant information in the dataset. The oversampling method, on the other hand, adds a minority class sample to an imbalanced dataset. The most common way is to directly copy the minority class sample so that a balanced amount of data is obtained between the minority class and the majority class. Even if the oversampling technique does not cause the loss or reduction of data information, there is also a possible risk that it can cause over-fitting.

Many oversampling techniques have been proposed by researchers, but the most common is the so-called The Synthetic Minority Oversampling Technique (SMOTE) which was first proposed by Chawla et al. [19]. The SMOTE algorithm is the most common oversampling method and the most popular approach for dealing with machine learning problems on imbalanced data [13]. Many researchers often use the SMOTE technique to balance data on two-class classification problems [20, 21], for example the study conducted by Khusi et al., where SMOTE was applied to a lung cancer dataset consisting of two classes [22]. The application of SMOTE to balance binary classification data related to fraud detection in the financial sector has been carried out by Hsin et al. [23]. Research on anomaly detection in industrial control systems conducted by Jiang et al. also applied SMOTE to balance the data for the purpose of classifying the two classes [24]. Guo et al. conducted a study to detect students' mental health by using multimodal education data pooling [25]. Obiedat et al. applied SMOTE to balance two classes of data in a research on customer satisfaction analysis [26]. The SMOTE process which is only carried out once cannot solve the problem of multi-class imbalanced data where the number of imbalanced data classes is more than one class, because with SMOTE once one class has become balanced, but the other classes remain a minority class. Although it is still rarely done, the use of SMOTE to balance multi-class data has been carried out by several researchers by adding certain strategies or techniques to the original SMOTE, for example by carrying out a weighting strategy as done by Deng et al. [27]. The difference between this study and previous studies is that it applies SMOTE repeatedly to balance data in multi-class classifications. The strategy applied is by repeating SMOTE as many as the existing minority class. After SMOTE is applied repeatedly and the data has become balanced, various classification algorithms will be applied to the balanced data. So the purpose of this study is to apply SMOTE repeatedly to the classification of multi-class imbalanced datasets. The classification algorithm used is neural network, k-NN, Nave-Bayes, and Random Forest. The selection of these four algorithms is based on the consideration that the four algorithms have different theoretical bases. Neural network imitates the workings of the human brain, k-NN uses the concept of nearest neighbourhood, Nave Bayes is based on statistics, while Random Forest is based on decision trees.

This paper contains four sub-chapters in the following order: introduction, research methods, results and analysis, and conclusions. The introduction contains the state of the art of this research. The research method contains the steps or stages of research. The results and analysis subsection contains the experimental results and analysis, and the last subsection is the conclusion.

2. RESEARCH METHOD

The research method was carried out following the steps shown in Figure 1 which included: preparing the dataset, the SMOTE process, determining the data separation method and its proportions, classifying the dataset, and calculating performance. Each step is explained in each subsequent sub-chapter.

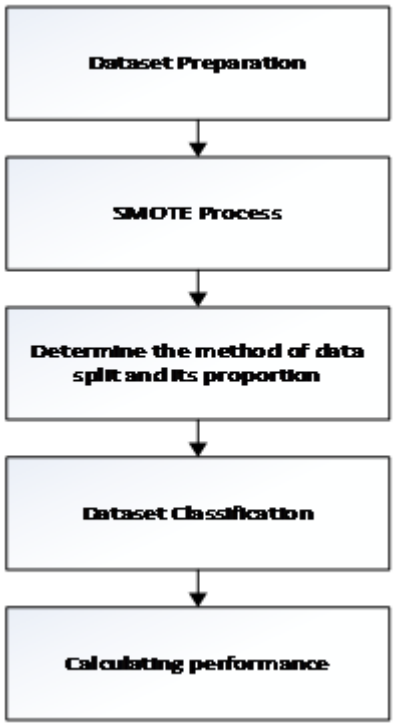


Figure 1. Research Steps

2.1. Dataset Preparation

The dataset used is taken from the UCI Machine Learning Repository which has a multi-class classification objective function. The dataset used is Glass Identification which consists of 214 data, has 10 input attributes, one of which is an attribute for id, while the output attribute is a classifier consisting of 7 classes, but one class of which has a frequency of 0. Frequency for each class shown in Table 1.

Table 1. Glass Classification dataset frequency per class

No	Class	Frequency
1	building_windows_float_processed	76
2	building_windows_non_float_processed	70
3	vehicle_windows_float_processed	29
4	vehicle_windows_non_float_processed	0
5	containers	17
6	tableware	13
7	headlamps	9
	Total	214

For class number 4, because it has a frequency of 0, it is not used in the SMOTE process or in the classification process, so the final dataset only has 6 classes. From this dataset, it can be seen that the frequency distribution for each class is less evenly distributed, there are classes that have relatively higher frequencies than other classes. Therefore, this dataset is a multi-class imbalanced dataset.

2.2. SMOTE Process

Usually the SMOTE process is carried out for datasets that have two classes, namely the majority class and the minority class. With the SMOTE process, previously the amount of data between the majority and minority class was not balanced to be balanced. To balance the multi-class dataset, the SMOTE process is carried out many times as many as the number of classes that have a smaller frequency than the class that has the largest frequency. For example, for the Glass Identification dataset, the SMOTE process is carried out 5 times. The data balancing algorithm uses the SMOTE technique proposed by Chawla et al. [19]. The principle of this algorithm is to increase the amount of data in the minority class so that the number is balanced with the majority class. The steps of the iterative SMOTE process can be seen in the flowchart in Figure 2.

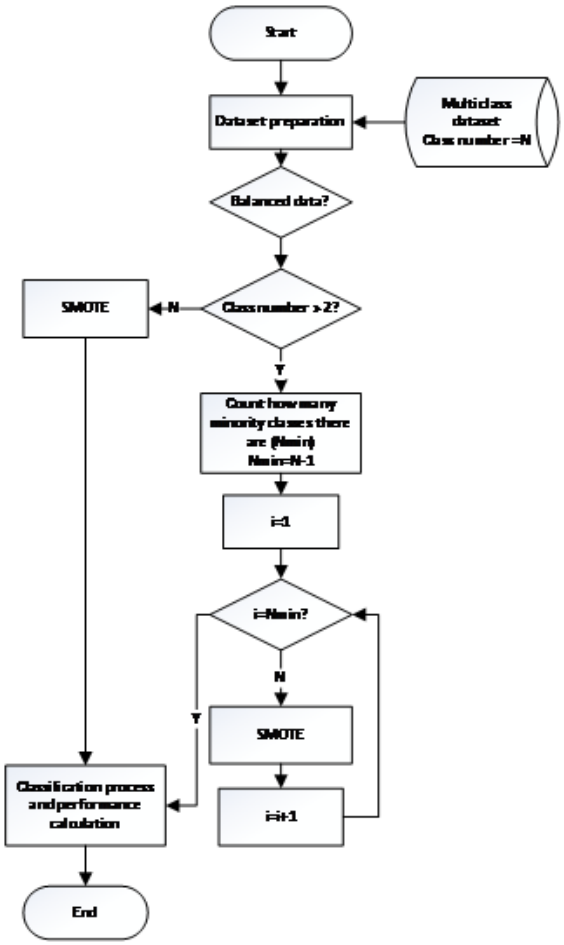


Figure 2. Repeated SMOTE Flowchart

2.3. Determining Data Split Methods and Their Proportions

The data split method used is cross-validation(CV), a statistical technique for evaluating and comparing learning algorithms by dividing the data into two parts: one is used to train the model and the other is used to validate or test the model [28, 29]. In cross-validation, the training and testing data must be exchanged sequentially so that each data point has a chance to be tested or validated. The basic form of cross-validation is often called k-fold cross-validation. In k-fold cross-validation, the data is partitioned into k equal-sized segments. Furthermore, the training and testing iterations are carried out in such a way that in each iteration one part of the data is used as test data while the remaining k-1 is used for training. In data mining and machine learning, 10-fold cross-validation is the most commonly used. An illustration of the 10-fold cross-validation can be seen in Table 2.

Table 2. Pembagian data untuk Training dan Testing

Experiment number	Dataset
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

From Table 2 it can be explained that by using 10-fold cross-validation, when running a machine learning process, it is as if the experiment was carried out 10 times, where in each experiment 10% of the data is for testing (blue box), and 90% of data for training. So the proportion of training data with testing is 90% versus 10% for each experiment. The accuracy of the 10-fold cross-validation method was obtained from the average accuracy value of 10 experiments. The advantage of using the cross-validation method is that all data can be used as training data and test data for different experiments.

2.4. Dataset Classification

There are four classification methods used in this study, namely:

1. Neural network
2. k-NN
3. Nave Bayes
4. Random Forest

Neural network is a model that imitate the workings of neurons in the human brain, where each neuron is interconnected to transmit information. This method is designed to solve problems that cannot be solved by conventional computing. Neural networks consist of a collection of neurons that interact with each other to solve cases such as classification and regression. In using this method, a neuron can have several inputs to process and only have one output which is the final result. The neural network structure consists of 3 layers, namely the input layer, the hidden layer or the layer between the input and output, and the last layer is the output layer [30]. The neural network architecture used is as presented in Figure 3, which consists of an input layer (X) which has 10 input attributes, using a hidden layer (h1) which has two neurons, and an output layer (Y) which will classify the output into seven classes.

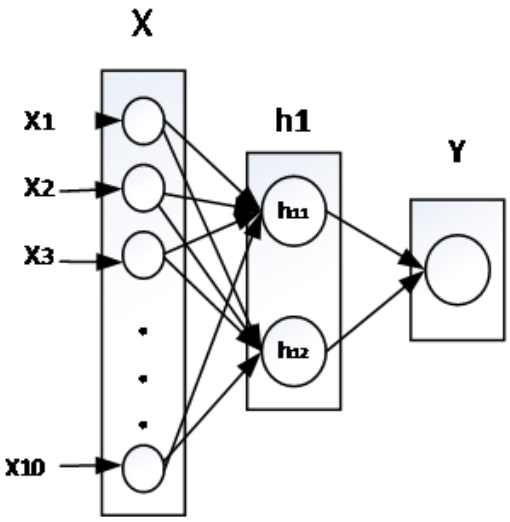


Figure 3. Neural Network Architecture Used

K-nearest neighbors (k-NN) is a technique to perform classification based on learning data, by taking the k nearest-neigbors [31]. The amount of data and the size of the dimensions of the data owned will affect the determination of the number of k neighbors. The proximity of a point to its neighbors can be calculated using the Euclidean distance (1).

$$(x, y) = \sqrt{\Sigma_{k=1}^{n}(a_k - b_k)^2} \tag{1}$$

a = point whose class is known, b = new point, and D = distance. D(x,y) is the distance between the known classification point (a) and the new point (b). The distance between the new point and the training data point will be calculated and the number of neighbors (k) is determined in the analysis.

Nave Bayes is a statistically based classifier applied to estimate probabilities for class membership. Bayes' theorem forms the basis of this method. This method uses the concept of probability with the assumption that the existing categories are independent or not dependent on other categories. This method predicts the probability of an event based on previous events. Based on the above equation, the value of the posterior or probability in class X with certain sample properties can be obtained from the probability of occurrence of class X before the sample or called prior, multiplied by the likelihood or probability of occurrence of the sample trait in class X, then divided by the probability of occurrence of the sample trait as a whole or evidence. The equation of the Bayes theorem is (2).

$$Posterior = \frac{prior \times likehood}{evidence} \tag{2}$$

Random Forest is a technique based on a collection of decision trees. This technique builds a decision tree starting from the root node to the leaf node by taking random attributes [32]. Method that uses a decision tree consisting of a root node, an internal node, and a leaf node [33]. The steps of Random Forest are as follows:

a. Determine the number of trees (k) selected from a total of m features, where k is less than m.
b. Take N samples of the dataset for each tree.
c. In each tree, take m subsets of predictors at random, where m< is the number of predictor variables.
d. Repeat the second and third steps until there are k trees.
e. Prediction results are obtained from the most votes from the classification results as many as trees.

The four methods used that have been described above have their respective parameters. The parameters used for each method are as follows:

1. Neural network uses a multilayer perceptron (MLP) architecture with one hidden layer with two neurons, cycles = 200, LR = 0.01, and momentum = 0.9
2. k-NN uses k=5 with the method of calculating distance using Euclidian distance
3. Naive Bayes, to avoid calculations that result in zero values using Laplace_correction
4. Random Forest, number of trees 100, criteria=gain ratio, maximum depth=10, voting strategy=confidence vote

2.5. Calculating Performance

The performance of each classifier method used is calculated based on the values of accuracy, sensitivity, and specificity. Accuracy is the percentage of the number of data that is correctly predicted to the total amount of data. Sensitivity is how reliably (expressed in percent) a model can correctly predict data that has a positive label. Specificity is a measure of the model's performance (in percent) in correctly predicting data that has a negative label. To get the performance of the classification by using SMOTE iteratively using a confusion matrix for multi-class classification. The formula for calculating each performance is presented in Table 3.

Table 3. Performance Measure Formula

Performance	Formula
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Sensitivity	(TP)/(TP+FN)
Specifisity	(TN)/(TN+FP)

TP= True positive, TN= True negative, FP= False positive, TN= True Negative

3. RESULT AND ANALYSIS

The results and analysis section presents the results of applying SMOTE to multi-class classification for the Glass Identification dataset. Based on Table 1, the datasets are grouped into 6 classes (classes with a frequency of 0 are not included), where the largest class frequency is 76. The other classes that have a smaller frequency are carried out in a stepwise SMOTE process, starting with the class that has the smallest frequency. The testing technique carried out in this study uses k-fold cross validation with the number of k = 10.

3.1. Classification without SMOTE

The original data from the dataset consisting of 214 data, was classified using the neural netwok machine learning method. The neural network architecture used is Multi-Layer Perceptron (MLP) which has a hidden layer with two neurons. The distribution of the data uses the 10-fold cross-validation method. After running the machine learning process using a neural network, the accuracy value is 61.69%, sensitivity is 44.38%, and specificity is 88.88%.

3.2. Classification with 1st SMOTE

In the 1st SMOTE process, the data balancing process is carried out for the class that has the smallest frequency = 9 (headlamp class) against the class that has the largest frequency = 76 (building_windows_float_processed class). After the 1st SMOTE process is carried out, the frequency for each class becomes as presented in Table 4.

Table 4. Dataset Frequency after 1st SMOTE

No	Class	Frequency	nth SMOTE
1	building_windows_float_processed	76	
2	building_windows_non_float_processed	70	
3	vehicle_windows_float_processed	29	
4	containers	17	
5	tableware	13	
6	headlamps	76	1st SMOTE
	Total	281	

So the headlamp class which originally had a frequency of 9 has a frequency of 76, so the number of datasets becomes 281. After the 1st SMOTE process is carried out, the dataset is then classified using a neural network and obtained an accuracy of 74.77%, sensitivity 62.67%, and specificity 92.87%.

3.3. Classification with 2nd SMOTE

In the 2nd SMOTE process, the data balancing process is carried out for the class that has the 2nd smallest frequency, namely the tableware class with a frequency of 13, against the class that has the largest frequency=76 (class building_windows_float_processed). After the 2nd SMOTE process is carried out, the frequency for each class becomes as shown in Table 5.

Table 5. Dataset Frequency After 2nd SMOTE

No	Class	Frequency	nth SMOTE
1	building_windows_float_processed	76	
2	building_windows_non_float_processed	70	
3	vehicle_windows_float_processed	29	
4	containers	17	
5	tableware	76	1st SMOTE
6	headlamps	76	2nd SMOTE
	Total	344	

So the tableware class which originally had a frequency of 13 became a frequency of 76, so the number of datasets became 344. After the 2nd SMOTE process was carried out, the dataset was then classified using a neural network and obtained an accuracy of 78.77%, a sensitivity of 68.96% and a specificity of 94.35%.

3.4. Classification with 3rd SMOTE, and 5th SMOTE

Using the same steps, through the 3rd SMOTE 4th SMOTE, and 5th SMOTE process, then the classes that have not been balanced data will be balanced against the class that has the greatest frequency. The final results obtained after the 5th SMOTE process are as presented in Table 6.

Table 6. Dataset Frequency After 5th SMOTE

No	Class	Initial Frequency	Final Frequency	nth SMOTE
1	building_windows_float_processed	76	76	
2	building_windows_non_float_processed	70	76	5th SMOTE
3	vehicle_windows_float_processed	29	76	4th SMOTE
4	containers	17	76	3rd SMOTE
5	tableware	13	76	2nd SMOTE
6	headlamps	9	76	1st SMOTE
	Total	214	380	

The results of the classification with the neural network after the 3rd SMOTE, 4th SMOTE, and 5th SMOTE resulted in the accuracy levels being: 73.97%, 78.00%, and 77.42%; sensitivity 74.60%, 77.63%, and 77.41%, respectively; and specificity 93.68%, 92.99%, and 92.91%, respectively. As a summary of the accuracy values obtained for each of the above stages are as presented in Table 7.

Table 7. Nilai Akurasi untuk SMOTE ke-n

No.	SMOTE ke-n	Accuracy	Sensitivity	Specificity
1	Without SMOTE	61.69%	44.38%	88.88%
2	1st SMOTE	74.77%	62.67%	92.87%
3	2nd SMOTE	**78.77%**	68.96%	**94.35%**
4	3rd SMOTE	73.97%	74.60%	93.68%
5	4th SMOTE	78.00%	**77.63%**	92.99%
6	5th SMOTE	77.42%	77.41%	92.91%

If depicted in graphical form, it can be seen in Figure 4. So for the Glass Identification dataset, the performance of the classifier using a neural network, the highest accuracy value is achieved when the SMOTE process is carried out twice (accuracy = 78.77%), while the lowest accuracy is when without processing without SMOTE (accuracy=61.69%). The highest sensitivity is 77.41% with four times SMOTE, while the lowest is 44.38 without SMOTE. The highest specificity was 94.35% with twice SMOTE, and the lowest was 88.88% without SMOTE.

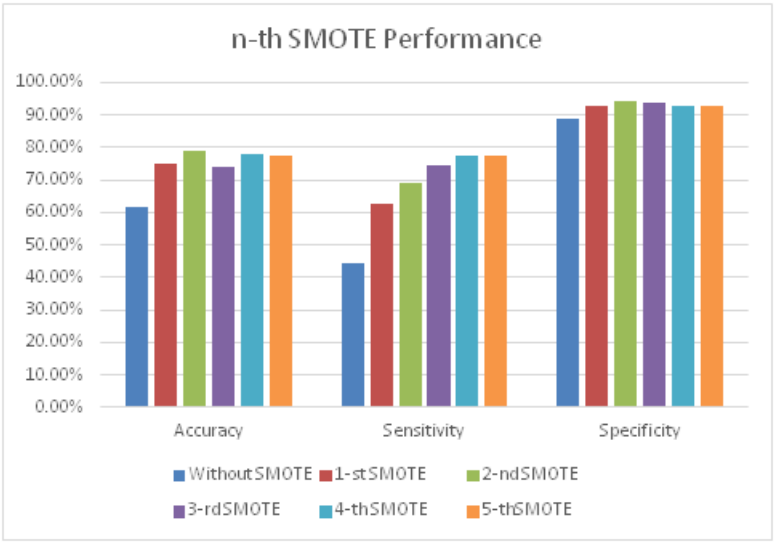


Figure 4. Performance for the nth SMOTE

3.5. Comparison of Classification Methods

In addition to using a neural network, in this study also perform a classification process using other methods, namely k-NN, Nave-Bayes, and Random Forest. The results of each nth SMOTE stage using several classifier methods can be seen in Table 8, Table 9, and Table 10 and in graphical form in Figure 5 for accuracy performance as an example.

Table 8. Comparison of Classification Method Accuracy

Methods	Without SMOTE	1st SMOTE	2nd SMOTE	3rd SMOTE	4th SMOTE	5th SMOTE
Neural Network	62.21%	74.77%	78.77%	73.97%	78.00%	77.42%
k-NN	69.16%	77.93%	81.70%	80.67%	83.33%	83.77%
Nave Bayes	44.33%	59.75%	63.09%	60.54%	65.33%	64.24%
Random Forest	72.42%	80.43%	82.28%	84.38%	85.78%	**86.17%**

Table 9. Comparison of Classification Method Sensitivity

Methods	Without SMOTE	1st SMOTE	2nd SMOTE	3rd SMOTE	4th SMOTE	5th SMOTE
Neural Network	44.38%	62.67%	68.96%	74.6%	77.63%	77.41%
k-NN	61%	67.82%	74.32%	79.82%	83.03%	83.77%
Nave Bayes	48.07%	54.05%	61.54%	62.68%	64.83%	64.26%
Random Forest	69.63%	71.62%	72.42%	83.92%	85.48%	**86.18%**

Table 10. Comparison of the Specificity of Classification Methods

Methods	Without SMOTE	1st SMOTE	2nd SMOTE	3rd SMOTE	4th SMOTE	5th SMOTE
Neural Network	88.88%	92.87%	94.35%	93.68%	92.99%	92.91%
k-NN	90.93%	93.85%	95.15%	95.14%	94.12%	94.24%
Nave Bayes	85.56%	90.18%	91.28%	90.90%	90.46%	90.28%
Random Forest	91.48%	94.24%	95.15%	**95.82%**	94.55%	94.63%

From Table 8, Table 9, and Table 10 and Figure 5, it can be seen that in general it can be said that the classification process for the Glass Identification dataset, using SMOTE has a higher level of accuracy, sensitivity, and specificity than without the SMOTE process.

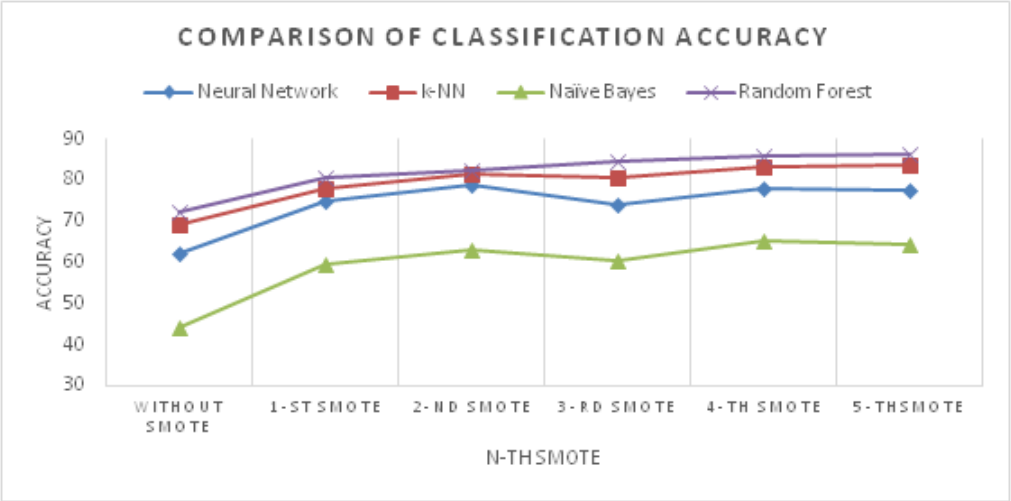


Figure 5. Comparison of the accuracy of classifier method

The best performance for each measure is achieved when the SMOTE process is performed three, four, or five times. This shows the performance of the classifier method for multi-class classification whose data is not balanced, it is necessary to carry out an iterative SMOTE process. The best level of accuracy of the four classifier methods used was achieved when SMOTE was performed more than once. For the neural network method, the best accuracy is achieved with two SMOTEs, Nave-Bayes after four SMOTEs, while for k-NN and Random Forest the best accuracy is obtained after five SMOTEs. For the neural network method, the best sensitivity is achieved with four SMOTEs, Nave-Bayes after four SMOTEs, while for k-NN and Random Forest the best accuracy is obtained after five SMOTEs. For the neural network method, the best specificity is achieved with four SMOTEs, Nave-Bayes after four SMOTEs, while for k-NN and Random Forest the best accuracy is obtained after five SMOTEs. So in general it can be said for the classification of multi-class datasets, to achieve the best accuracy, sensitivity, and specificity, the SMOTE process is required more than once. Based on the level of accuracy per SMOTE stage, the use of Random Forest for the classification of multi-class Glass Identification datasets provides the best performance. The good performance of the Random Forest algorithm is in accordance with research conducted by Santosa et al. [33]. Random Forest is a reliable classification algorithm for the case of unbalanced data [34, 35]. In a previous study conducted by Szepannek and Perry [36], the performance measure of accuracy with the Random Forest method reached 79.1%, while that conducted by Aldayel [37] using the k-NN method obtained the best performance for accuracy = 80.37%. , and sensitivity = 80.4%. Mathur and Surana [38] conducted a study with the best performance for the Random Forest method with accuracy = 79.62%, and using the k-NN method, accuracy = 74.07%. So the results obtained in this study have a better performance than the three studies.

4. CONCLUSION

Imbalanced data presents difficulties for many classifier algorithms. Oversampling of training data to make it more even is an effective strategy to solve this problem at the data processing level, one of which is through the SMOTE method. Generally, SMOTE is widely applied to the classification of two classes so that the SMOTE process is only applied once. In this study, SMOTE was applied to multi-class classification problems. The strategy is to do an iterative SMOTE process. Experimental results with various classification methods show that the repeated SMOTE process provides better accuracy, sensitivity, and specificity than applying the classifier method without SMOTE or only doing SMOTE once. The best performance for each measure of Glass Identification dataset classification with repeated SMOTE is: accuracy = 86.17%, sensitivity = 86.18%, and specificity = 95.82% where the classification technique used is Random Forest. For further research, the properties of each classifier need to be analyzed related to how many SMOTE iterations are needed to obtain the best level of accuracy and also research needs to be carried out to be applied to other classifier algorithms.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AUTHOR CONTIBUTION

FUNDING STATEMENT

COMPETING INTEREST

REFERENCES

[1] F. Bao, Y. Wu, Z. Li, Y. Li, L. Liu, and G. Chen, "Effect Improved for High-Dimensional and Unbalanced Data Anomaly Detection Model Based on KNN-SMOTE-LSTM," *Complexity*, vol. 2020, pp. 1–17, 2020.

[2] J. Luo, L. Zhu, Q. Li, D. Liu, and M. Chen, "Imbalanced Fault Diagnosis of Rotating Machinery Based on Deep Generative Adversarial Networks with Gradient Penalty," *Processes*, vol. 9, pp. 1–13, 2021.

[3] Y. Fan, X. Cui, H. Han, and H. Lu, "Chiller Fault Diagnosis with Field Sensors Using The Technology of Imbalanced Data," *Applied Thermal Engineering*, vol. 159, p. 113933, aug 2019.

[4] H. Hairan, K. E. Saputro, and S. Fadli, "K-means-SMOTE for Handling Class Imbalance in The Classification of Diabetes with C4.5, SVM, and Naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, apr 2020.

[5] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote dan K-Nearest Neighbor," *Information System Development*, vol. 3, no. 1, pp. 44–49, 2018.

[6] M. Koziarski, "Potential Anchoring for Imbalanced Data Classification," *Pattern Recognition*, vol. 120, p. 108114, dec 2021.

[7] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64 606–64 628, 2021.

[8] M. Mukherjee and M. Khushi, "SMOTE-ENC : A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features," *Applied System Innovation*, vol. 4, no. 12, pp. 1–12, 2021.

[9] J. Liu, "Importance-SMOTE: A Synthetic Minority Oversampling Method for Noisy Imbalanced Data," *Soft Computing*, vol. 26, no. 2, pp. 1141–1163, 2022.

[10] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on Expansion and Classification of Imbalanced Data Based on SMOTE Algorithm," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[11] A. Guezzaz, Y. Asimi, M. Azrour, and A. Asimi, "Mathematical Validation of Proposed Machine Learning Classifier for Heterogeneous Traffic and Anomaly Detection," *Big Data Mining and Analytics*, vol. 4, no. 1, pp. 18–24, mar 2021.

[12] L. Huang, Q. Fu, M. He, D. Jiang, and Z. Hao, "Detection Algorithm of Safety Helmet Wearing Based on Deep Learning," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 13, pp. 1–14, jul 2021.

[13] A. Fernández, S. García, M. Galar, and R. C. Prati, *Learning From Imbalanced Data Sets*. Switzerland AG: Springer, 2018.

[14] D. Veganzones and E. Séverin, "An Investigation of Bankruptcy Prediction in Imbalanced Datasets," *Decision Support Systems*, vol. 112, pp. 111–124, aug 2018.

[15] M. Pirizadeh, N. Alemohammad, M. Manthouri, and M. Pirizadeh, "A New Machine Learning Ensemble Model for Class Imbalance Problem of Screening Enhanced Oil Recovery Methods," *Journal of Petroleum Science and Engineering*, vol. 198, p. 108214, mar 2021.

[16] S. V. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE, 2018, pp. 1–11.

[17] L. Yu, R. Zhou, L. Tang, and R. Chen, "A DBN-Based Resampling SVM Ensemble Learning Paradigm for Credit Classification with Imbalanced Data," *Applied Soft Computing*, vol. 69, pp. 192–202, aug 2018.

[18] A. S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3853–3864, 2021.

[19] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.

[20] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A Cluster-based Oversampling Algorithm Combining SMOTE and K-means for Imbalanced Medical Data," *Information Sciences*, vol. 572, no. 5, pp. 574–589, sep 2021.

[21] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with Cost Sensitive Classification Algorithm for Imbalanced Data in Medical Diagnosis," *Computers and Industrial Engineering*, vol. 140, p. 106266, 2020.

[22] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[23] Y.-y. Hsin, T.-s. Dai, Y.-w. Ti, M.-c. Huang, T.-h. Chiang, and L.-c. Liu, "Feature Engineering and Resampling Strategies for Fund Transfer Fraud with Limited Transaction Data and A Time-Inhomogeneous Modi Operandi," *IEEE Access*, vol. 10, no. August, pp. 86 101–86 116, 2022.

[24] J.-r. Jiang and Y.-t. Chen, "Industrial Control System Anomaly Detection and Classification Based on Network Traffic," *IEEE Access*, vol. 10, pp. 41 874–41 888, 2022.

[25] T. Guo, W. Zhao, M. Alrashoud, A. Tolba, S. Firmin, and F. Xia, "Multimodal Educational Data Fusion for Students' Mental Health Detection," *IEEE Access*, vol. 10, no. May, pp. 70 370–70 382, 2022.

[26] R. Obiedat, R. Qaddoura, A. M. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, and H. Faris, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22 260–22 273, 2022.

[27] M. Deng, Y. Guo, C. Wang, and F. Wu, "An Oversampling Method for Multi-Class Imbalanced Data Based on Composite Weights," *PLOS ONE*, vol. 16, no. 11, p. e0259227, nov 2021.

[28] N. Darapureddy, N. Karatapu, and T. K. Battula, "Research of Machine Learning Algorithms Using K-fold Cross Validation," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6S, pp. 215–218, sep 2019.

[29] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, dec 2021.

[30] M. L. Suliztia and A. Fauzan, "Comparing Naive Bayes , K-Nearest Neighbor , and Neural Network Classification Methods of Seat Load Factor in Lombok Outbound Flights," *Jurnal Matematika, Statistika & Komputasi*, vol. 16, no. 2, pp. 187–198, 2020.

[31] A. Naimi, J. Deng, and S. Member, "Fault Detection and Isolation of a Pressurized Water Reactor Based on Neural Network and K-Nearest Neighbor," *IEEE Access*, vol. 10, pp. 17 113–17 121, 2022.

[32] P. R. Sihombing and I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, pp. 417–426, may 2021.

[33] N. Santoso, W. Wibowo, and H. Hikmawati, "Integration of Synthetic Minority Oversampling Technique for Imbalanced Class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, p. 102, jan 2019.

[34] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased Random Forest for Dealing with The Class Imbalance Problem," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 2163–2172, jul 2019.

[35] A. S. More, D. P. Rana, and I. Agarwal, "Random Forest Classifier Approach for Imbalanced Big Data Classification for Smart City Application Domains," in *Proceedings of International Conference on Computational Intelligence & IoT (ICCIIoT)*, 2018, pp. 260–266.

[36] and T. Perry G. Szepannek, "Explaining Artificial Intelligence with Care," *Künstl Intell*, pp. 1–10, 2022.

[37] M. S. Aldayel, "K-Nearest Neighbor Classification for Glass Identification Problem," in *International Conference on Computer Systems and Industrial Informatics*, 2012, pp. 1–5.

[38] H. Mathur and A. Surana, "Glass Classification Based on Machine Learning Algorithms," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 11, pp. 139–142, 2020.