Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 9, No. 4, December 2023, pp. 973-982
ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i4.26881

Gaussian Based-SMOTE Method for Handling Imbalanced Small Datasets

Muhammad Misdram[1,2], Edi Noersasongko[1], Purwanto[1], Muljono[1], Fandi Yulian Pamuji[2]
[1]Faculty of Computer Science, Universitas Dian Nuswantoro. Semarang 51031 Indonesia
[2]Faculty of Information Technology Merdeka University, Pasuruan 67129 Indonesia

ARTICLE INFO

Article history:

Received August 10, 2023
Revised September 22, 2023
Published October 16, 2023

Keywords:

SMOTE;
Gaussian;
Imbalance;
Oversampling;
Undersampling

ABSTRACT

The problem of dataset imbalance needs special handling, because it often creates obstacles to the classification process. A very important problem in classification is to overcome a decrease in classification performance. There have been many published researches on the topic of overcoming dataset imbalances, but the results are still unsatisfactory. This is proven by the results of the average accuracy increase which is still not significant. There are several common methods that can be used to deal with dataset imbalances. For example, oversampling, undersampling, Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE, Adasyn, Cluster-SMOTE methods. These methods in testing the results of the classification accuracy average are still relatively low. In this research the selected dataset is a medical dataset which is classified as a small dataset of less than 200 records. The proposed method is Gaussian Based-SMOTE which is expected to work in a normal distribution and can determine excess samples for minority classes. The Gaussian Based-SMOTE method is a contribution of this research and can produce better accuracy than the previous research. The way the Gaussian Based-SMOTE method works is to start by determining the random location of synthesis candidates, determining the Gaussian distribution. The results of these two methods are substituted to produce perfect synthetic values. Generated synthetic values are combined with SMOTE sampling of the majority data from the training data, produce balanced data. The result of the balanced data classification trial from the influence of the Gaussian Based SMOTE result in a significant increase in accuracy values of 3% on average.

Corresponding Author:

Muljono, Faculty of Computer Science, Dian Nuswantoro University Semarang 51031 Indonesia
E-mail: muljono@dsn.dinus.ac.id

1. INTRODUCTION

The standard learning method assumes that the training estimates are a balanced dataset [1]. Almost all datasets have the same performance importance in their classes. Even in real-life scenarios, balanced datasets often experience higher errors in the classification process [2]. Dataset imbalance can arise due to the same unrepresented class, resulting in a skewed number of classes due to differences in the substantial number of data classes. This will affect the data quality of the classification results [3]. The imbalance of datasets has been the subject of extensive research and the application of numerous approaches, yet, the robustness of the method findings is still inadequate. Concrete evidence is that the classification results have not resulted in a significant increase in accuracy. Many methods exist to overcome data imbalances, but the results are still unsatisfactory. There are many data imbalance problems, so new research contributions and innovations are still needed to apply more methods and methods.

There are several cases of data imbalance impacting real domains, such as the assessment of cancer malignancy, fraud detection, and behavior analysis. The problem of imbalanced data classification will remain an active object of research.

Many instances of data imbalance affect real domains, as these examples will show. For example, the evaluation of the malignancy of cancer [4], the identification of fraud [5], and the analysis of behavior [6], the diabetes prediction [7]. The issue of unbalanced data classification will continue to be a topic of ongoing research [8], [9]. When it comes to addressing data imbalances, a significant body of published work has been put out as potential solutions. An environment-based strategy that is based on the Synthetic Minority Oversampling Technique (SMOTE) [10] is one of the most common approaches that is taken to address data imbalances. Synthetic minority oversampling, often known as the SMOTE method, is a methodology that can provide a direct solution to the problem of data imbalance [11], [10]. Its benefits include the capacity to construct synthetic samples until the requisite number of samples has been created, guided by a combination of the k-nearest neighbour (k-NN) method and a uniform probability distribution. Regarding the flaw, there is no consideration of neighbouring information from minority class samples. This is a significant limitation. While this is going on, synthetic samples are experiencing an overgeneration problem, which does not help in the fight against class imbalance. A number of different SMOTE approaches, such as Borderline-SMOTE [12], Safe-level SMOTE [13], DBSMOTE [14], EFN-SMOTE [15] and Cluster-SMOTE [16], can be utilized to address the inadequacies of the original method. But when it is known that the method has created instances of synthetic values with linear interpolation, one may see that the weakness of the synthetic data space is a bounded region. This can be detected when one knows that the method has made examples of synthetic values with linear interpolation [17]. It can go beyond the data that has been provided because there is insufficient room. The method that is based on distance does not take into account the statistical traits that are associated with the minority class. Depending on how the distance is defined or the parameters established for the sampling procedure, it might be challenging to locate and identify outliers. As a result of the distance-centered strategy, which does not evaluate the statistical characteristics of the minority class, it will be difficult to spot outliers, and overgeneration will not be dealt with appropriately.

It is important to classify the dataset to obtain the outcomes of the accuracy tests run on it. The process of machine learning includes classification techniques. Several other overarching categories are recognized for their excellent performance. One of these is called the Support Vector Machine (SVM), and it has a good level of accuracy most of the time, but it is not ideal for use with imbalanced data [18]. The difficulty when applying the SVM approach to large cases and two-class situations is another of the method's shortcomings [19], [20]. The Decision Tree, or D-Tree, method makes decision-making much simpler; however, this method frequently results in overlapping [21], [22] . According to two studies [20], [21] that produced high accuracy when used to predict heart disease and classify sarcastic tweets, the random forest approach is a well-known random forest method that produces high accuracy. This is shown by the random forest method creating high [23], [24]. The technique involved in the NB method is less complicated and reliably generates high-accuracy values [25], [26], [27]. The k-nearest neighbour (k-NN) method has the benefit of being able to handle noisy training data effectively and has the benefit of not require the determination of the neighbour value (k) [28], [29], [30], [31].

In this research, the Gaussian Based-SMOTE method is proposed. This method intends to improve the performance of the previously used method, which is important because there is a high amount of unbalanced data. The Gaussian-Based SMOTE approach operates on the same principle as the SMOTE method, which involves taking excessive samples for the minority class. The next step is to mix the samples produced by the SMOTE method with the synthetic data produced by the Gaussian distribution to build a data balance. The Gaussian distribution is responsible for producing the synthetic data. So, as a contribution of this research is Gaussian Based-SMOTE method and It expected can increase the accuracy value in the small unbalanced datasets classification. The Naeve Bayes, k-NN, SVM, D-Tree, and RF classification algorithms are tested. The utilized dataset is a tiny public dataset with fewer than 200 records obtained from the UCI repository.

2. LITERATURE REVIEW

2.1. Related research

Previous research utilized an approach known as the Elbow Fuzzy Noice filtering SMOTE (EFN-SMOTE) method [15]. The dataset is partitioned into clusters for this method to function properly; the number of clusters is established using a technique known as the Elbow method. Following the application of each noise filtering to each cluster comes the employment of SMOTE, which generates a new minority instance for each cluster by basing it on the majority that is closest to it. An Artificial Neural Network (ANN) was used to classify based on the EFN-SMOTE trial's findings. The results that were obtained are as follows: accuracy equal to 0.999, precision equal to 0.998, sensitivity equal to 0.999, specificity equal to 0.998, F-measure equal

to 0.999, and G-Mean equal to 0.999 [15]. When combined with an Artificial Neural Network (ANN), the EFN-SMOTE approach is an effective classification tool.

The problem of unevenly distributed data in the prior research's outputs was solved using the hybrid technique [32], [33]. The decision tree method, k-nearest neighbour, and discriminant analysis utilized the classification approaches. In comparison to the previous strategies for addressing categorization problems, they result in significantly more improvement. The research paper "Gaussian-Based SMOTE Method for Solving Skewed Class Distributions" indicated that the first study stage tests the basic SMOTE approach. This means the initial data is processed using the SMOTE method using k neighbours with the k-NN method. The second approach uses synthetic data generated through randomization and the SMOTE, Borderline-SMOTE, and safe-level-SMOTE analysis methodologies. Third, the data are processed using a new SMOTE method known as Gaussian-based SMOTE. This method is predicted to produce a new color regarding the total amount of bogus data formed. The trial findings produce an accuracy of 87.30% when using real data, 89.11% when using artificial data generated by SMOTE, and 90.13% when using an artificial dataset generated by Gaussian-SMOTE [34], [35].

In general, altering the ratio of the imbalance ratio between the two classes that are in the minority and the class that is in the majority is the approach to solving the problem of the imbalance [36], [37]. Based on this explanation, several different sample strategies have been implemented to modify the characters based on the imbalanced data distribution. These sampling strategies include random oversampling and undersampling and the creation of synthetic minorities based on sampling methods [34], [38].

2.2. SMOTE

Utilizing k-nearest neighbours and a uniform probability distribution, the SMOTE approach has been reported to generate synthetic data. This is accomplished by using the method. The following is an explanation of how the SMOTE method works (Fig. 1). In the beginning, the method successfully partitioned the data given to the majority class from the data given to the minority class.

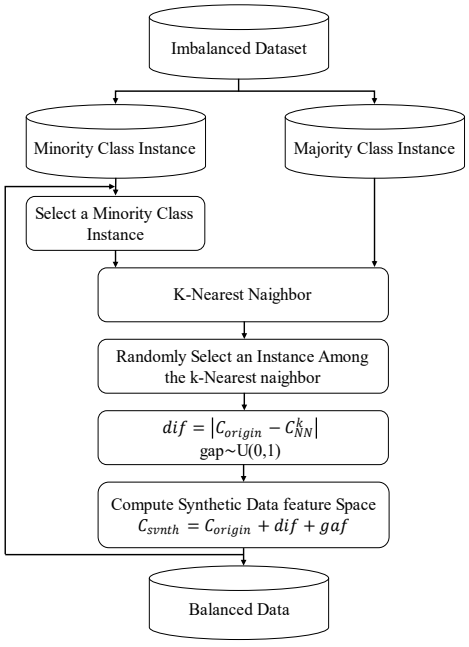

Fig. 1. SMOTE Method Flow

The following step will apply the k-Nearest Neighbour (k-NN) algorithm, resulting in each minority class data having k neighbours. When constructing synthetic samples, each minority sample is assigned a nearest neighbour from the set of k-nearest neighbours based on a random selection process. In this way, the synthetic samples are created. After that, the [34].

$$dif = |C_{origin} - C_{NN}^{k}| \quad (1)$$

C_{origin} = minority class data, C_{NN}^{k} = one of the k-nearest neighbors of minority class data by selecting random values from a uniform probability distribution. After that, the random value of the uniform probability

distribution is multiplied by the value of the variation to make the unpredictability a little bit more unpredictable. The following equation can be used to estimate the number of synthetic samples that can be produced.

$$dC_{synth} = C_{origin} + |C_{origin} - C_{NN}^{k}| x\ P_{unifor} \quad (2)$$

where P_{unifor} is random values from a uniform probability distribution. The procedure will keep cycling through the sequence described above until it satisfies the termination requirements, which includes the required total number of samples.

3. RESEARCH METHODOLOGY

The phases of the research methodology process that are carried out begin with the stages of the dataset collection process and continue to the stages of the data imputation process, the stages of the data imbalance process, and the stages of the classification process, respectively. The subsequent step is to generate output from the anticipated degree of precision. A few steps can be broken down like this in this research.

3.1. Data Collection

The datasets that were obtained were from the repository at UCI. Each collected dataset dealt with medical topics such as hepatitis, immunotherapy, and echocardiograms. The three selected datasets were considered small, each containing fewer than 200 records, and an imbalance in the data was discovered [39], [40].

3.2. Preprocessing

The step to get the dataset ready for testing is called preprocessing, and it's important. Datasets that are the product of excavation almost always have flaws; for instance, if an empty record is discovered, the dataset must be treated before it can be considered complete. Imputation is the most effective method for enhancing the dataset to keep it in its ideal state. In this study, the imputation method was the k-NN approach; however, there are many different methods of imputation from which one can select [41].

3.3. Gaussian Based-SMOTE

Random numbers with a uniform distribution have been used in the SMOTE, Borderline-SMOTE, Safe-level SMOTE, and DBSMOTE procedures and other approaches for producing synthetic data. However, it is also possible to forecast the production of more than one set of synthetic data. It is frequently chosen throughout the method selection process in particular minority class data and the data of its nearest neighbours. The placement of synthetic data on the same line has a high chance, and as a result, this can lead to significant overgeneration issues.

The Gaussian Based-Smote approach is presented as a solution to the issues discovered throughout this research. To generate random points connected to the minority class and take into consideration overgeneration, the goal of this method is to make it as likely as possible that it will operate based on a normal distribution. Taking excessively large samples from underrepresented groups is the fundamental idea behind this method, built on the SMOTE method's foundation. When the minority sample includes discrepancies between the minority class data and the data of the nearest neighbours, which may be determined randomly, the case is known as "where the minority sample contains." (1) displays the equation for your perusal. In addition, the Gaussian-based SMOTE approach will randomly forecast the position of synthetic candidates by selecting a number between 0 and the difference value, as demonstrated by the following equation. This number will be based on the difference value.

$$dGap = U(0, dif) \quad (3)$$

The following step involves taking another number from the Gaussian distribution, depicted in (4), and doing so by heuristically selecting the parameter δ.

$$d\text{Range} = \text{N (gap } \delta) \quad (4)$$

In conclusion, it is possible to construct synthetic data, which is indicated in (5), by deriving the derived parameters from (3) and (4). This may be done by following the steps shown in the previous sentence.

$$Csynthetic = Corigin + dif\ x\ range \quad (5)$$

at the classification stage, K-Fold Validation has been carried out, the purpose of which is dividing the training dataset into segments of the same or close size. The next stage is repetition of training data and validating.

Each repetition of different data is needed for validation, so that if it is applied to different training data on the same dataset, it will get the most effective and accurate results.

The following is a flowchart that may be used to represent the sequence of explanations about Gaussian Based SMOTE that can be found in this research (Fig. 2).

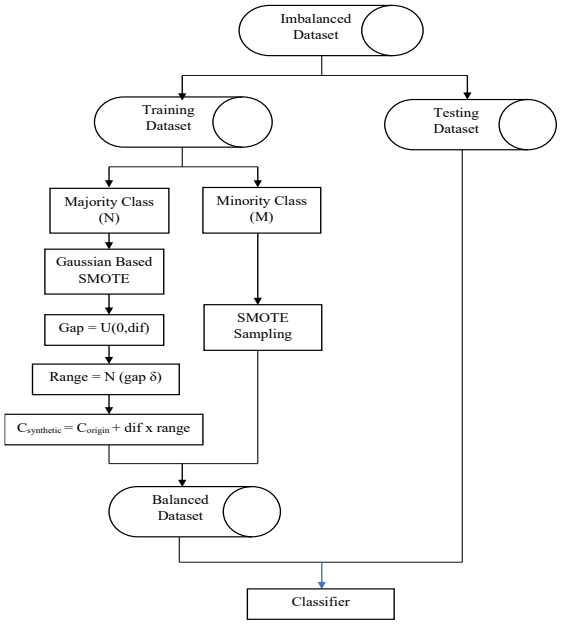

Fig. 2. Flowchart of the Gaussian Based-SMOTE method

Based on the the reseach flow with basic method of SMOTE in Fig. 1, It has been developed by combining the Gaussian statistic model or Gaussian Distribution to form the Gaussian Based-SMOTE method. The way of Gaussian Based-SMOTE works is start from an unbalanced datasets between training data and testing data. From the training dataset, the majority class and minority class are formed. These two classes will be manipulated with SMOTE method and Gaussian Based-SMOTE so that a new balanced dataset will be formed, while compared with the original testing datased, a balanced will be formed. The flow diagram can be drawn in Fig. 3.

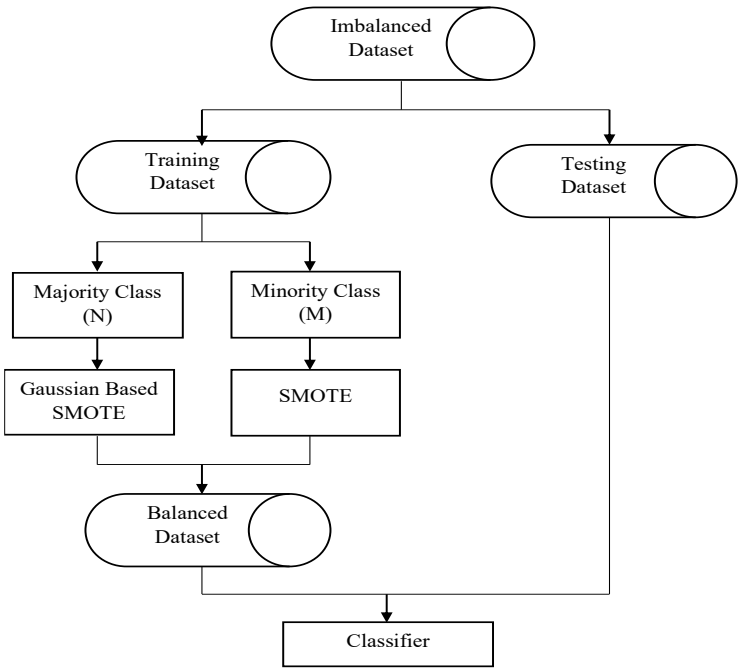

Fig. 3. Gaussian Based-SMOTE and SMOTE method Flowchart

4. RESULTS AND DISCUSSION

This research involved 3 selected datasets, all of which were medical datasets among others, namely hepatitis, echocardiogram, and immunotherapy which were taken from the UCI repository. The dataset can be described in Table 1. The description of the dataset is as follows, Hepatitis has 2 classes that represent death "Die" and life "live", and the Echordiogram dataset has 2 classes that represent death "Aliev" and life "Dead". Two groups, "No" and "Yes," are represented in the Immunotherapy dataset. Their goal is to assist treatment, and if a positive "Yes" is treated, treatment can be continued; if a negative "No," treatment can be abandoned.

Table 1. Dataset Description

Dataset	No. of Record	No. of Features	Imbalanced Ratio	Minority Class	Majority Class	Number of Classes
Hepatitis	155	20	80 : 20	"Die"	"Live"	2
Echordiogram	131	13	70 : 30	"Alive"	"Dead"	2
Immunotherapy	90	8	80 : 20	"No"	"Yes"	2

Fig. 4 compares the state of the initial unbalanced dataset and the state of the balanced dataset after being subjected to the influence of the Gaussian Based-SMOTE approach. The original dataset is imbalanced. For example, in the hepatitis dataset, the majority data is 124, while the minority data is 31. In the echocardiogram dataset, the majority data is 107, while the minority data is 24, and in the immunotherapy dataset, the majority data is 71, while the minority data is 19. The Gaussian Based-SMOTE approach is implemented, and the result is balanced, as shown in Fig. 3. This was done to overcome the data imbalances that were there.

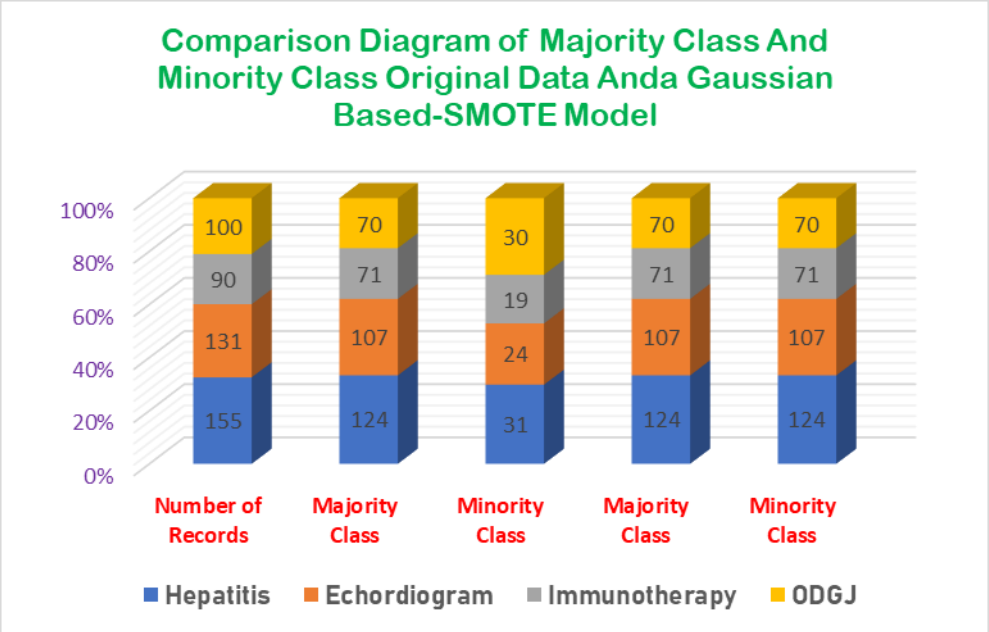

Fig. 4. Diagram of Comparison of Majority and Minority Actual data and the Gaussian Based SMOTE Model

The results of the tests for categorizing the dataset are presented in Table 2, which uses the computer language Python. This research test categorized three different datasets utilizing five different classification strategies. The datasets examined are imbalanced and balanced varieties, both of which are the products of using the Gaussian Based-SMOTE approach. The five different categorization approaches that were examined resulted in an improvement in the accuracy of each dataset. Table 2 makes it much more evident that the echocardiogram dataset has the highest unbalanced dataset accuracy value, which is 86%. This can be seen more clearly in the table. In contrast, the Echordiogram dataset has a balanced distribution, which allows for its best value of accuracy to reach 91%.

To make it easier to recognize dataset characters, the best way is to describe the dataset as listed in Table 1. The data description describes the data characters, for example, the number of records, attributes, data imbalance ratios, and minority and majority classes. In this research, the concern is the imbalance of majority and minority data. The dataset imbalance will affect the accuracy value of the data classification. The Gaussian Based-SMOTE method is used to overcome the dataset's imbalance in this research. The ways of the Gaussian

Based-SMOTE method works can be seen in Figure 2**.** The process starts with minority data or training data, and then a sampling process is carried out to become a minority and a majority. The Gaussian Based-SMOTE method generated the Synthetic data combined with SMOTE sampling to form a balanced dataset. The results of the classification test can be seen in Table 2**.**

Table 2. Accuracy value of dataset test results

Accuracy Value	Classification Method	Dataset Name		
		Hepatitis	Echordiogram	Immunotherapy
Imbalanced Dataset	NB	73	76	70
	SVM	**81**	79	**81**
	D-Tree	68	83	77
	k-NN	72	66	75
	RF	80	**86**	80
Balanced Dataset (Gaussian Based-SMOTE Effect)	NB	**84**	**91**	73
	SVM	81	66	77
	D-Tree	74	77	73
	k-NN	72	71	63
	RF	80	88	**82**

Imbalanced dataset classification results in the highest accuracy value of 86% in the Echordiogram dataset. In comparison, the classification of a balanced dataset produces the highest accuracy value of 91% in the Echordiogram dataset. In Table 3, the results of the imbalanced and balanced datasets' accuracy are influenced by the Gaussian-Based-SMOTE method. The hepatitis dataset has an increase in accuracy of 81% to 84% and an increase of 3%. The increase in the Echordiogram dataset from 86% to 91% increased accuracy by 5%. The Immunotherapy dataset has increased in accuracy from 81% to 82%, so there has been an increase in accuracy of 1%. The increase in the average accuracy value of imbalanced datasets and balanced datasets is 3%. This can be seen in Table 3. The accuracy value for the imbalanced dataset is 82.5%, while the accuracy value for the balanced dataset is 85.3%. Based on research conducted by [34], It has been tested with a benchmark dataset. The classification results of the original unbalanced dataset produced the highest accuracy of 87,30%. After balancing the dataset using Gaussian Based-SMOTE method, the accuracy increase to be 90,13%. Compared with current research by applying the same method, the accuracy maximum is 91%, there is difference in increase of 0.87%.

Table 3. The Increase in the results of the accuracy value on the Dataset

DataSet name	**Imbalanced Dataset**	**Balanced Dataset (Gaussian Based-SMOTE Effect)**	**Improved accuracy**
	Accuracy	**Accuracy**	
Hepatitis	81%	84%	3%
Echordiogram	86%	91%	5%
Immunotherapy	81%	82%	1%
Average Improved Accuracy	82.7%	85.3%	3%

5. CONCLUSION

Based on the results of the research above, it can be concluded that handling imbalanced datasets is really needed, so that the accuracy value will be better. The results of the classification tests of the 3 datasets have resulted in accuracy values for both imbalanced datasets and balanced datasets from the application of the Gaussian Based-SMOTE method. In this research, there was a significant increase in the accuracy value of the balanced dataset resulting from the application of the Gaussian-Based-SMOTE method. The resulting average accuracy value increases by 3%. The hypothesis that can be concluded is that the application of the Gaussian Based-SMOTE method to imbalanced datasets greatly affects the increase in accuracy values. This can be proven that the classification of imbalanced datasets has a smaller accuracy value compared to the classification of balanced datasets. In this research, this test is applied to a small dataset, while other tests need to be applied to a large dataset.

REFERENCES

[1] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," *Pattern Recognit.*, vol. 102, p. 107262, 2020, https://doi.org/10.1016/j.patcog.2020.107262.
[2] C. Eyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern*

Recognit., vol. 48, no. 5, pp. 1653–1672, 2015, https://doi.org/10.1016/j.patcog.2014.10.032.
[3] B. Prasetiyo, Alamsyah, M. A. Muslim, and N. Baroroh, "Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, 2021, https://doi.org/10.1088/1742-6596/1918/4/042002.
[4] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Appl. Soft Comput. J.*, vol. 38, pp. 714–726, 2016, https://doi.org/10.1016/j.asoc.2015.08.060.
[5] E. Ramentol *et al.*, "Fuzzy-rough imbalanced learning for the diagnosis of High Voltage Circuit Breaker maintenance: The SMOTE-FRST-2T algorithm," *Eng. Appl. Artif. Intell.*, vol. 48, pp. 134–139, 2016, https://doi.org/10.1016/j.engappai.2015.10.009.
[6] A. Azaria, A. Richardson, S. Kraus, and V. S. Subrahmanian, "Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data," *IEEE Trans. Comput. Soc. Syst.*, vol. 1, no. 2, pp. 135–155, 2014, https://doi.org/10.1109/TCSS.2014.2377811.
[7] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artif. Intell. Med.*, vol. 104, no. January, p. 101815, 2020, https://doi.org/10.1016/j.artmed.2020.101815.
[8] A. Fernández, C. J. Carmona, M. José Del Jesus, and F. Herrera, "A pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets," *Int. J. Neural Syst.*, vol. 27, no. 6, pp. 1–21, 2017, https://doi.org/10.1142/S0129065717500289.
[9] M. Lango and J. Stefanowski, "Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data," *J. Intell. Inf. Syst.*, vol. 50, no. 1, pp. 97–127, 2018, https://doi.org/10.1007/s10844-017-0446-7.
[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002, [Online]. Available: https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.
[11] S. Park and H. Park, "Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic," *Computing*, vol. 103, no. 3, pp. 401–424, 2021, https://doi.org/10.1007/s00607-020-00854-1.
[12] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *Lect. Notes Comput. Sci.*, vol. 3644, pp. 878–887, 2005, https://doi.org/10.1007/11538059_91.
[13] A. Mudinas *et al.*, "LOF: Identifying Density-Based Local Outliers," *Mach. Learn. Appl. An Int. J.*, vol. 9, no. 2, pp. 4–23, 2020, [Online]. Available: https://doi.org/10.1145/342009.335388.
[14] C. Bunkhumpornpat and K. Sinapiromsaran, "DBMUTE: density-based majority under-sampling technique," *Knowl. Inf. Syst.*, vol. 50, no. 3, pp. 827–850, 2017, https://doi.org/10.1007/s10115-016-0957-5.
[15] B. Kasasbeh and H. Ahmad, "EFN-SMOTE — An improved unbalanced data set oversampling based on fuzzy C-means for Credit Cards Fraud detection data set oversampling based on fuzzy," *Research Square,* pp. 0–16, 2022, https://doi.org/10.21203/rs.3.rs-2067504/v1.
[16] D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," *IEEE Int. Conf. Granul. Comput.*, pp. 732–737, 2006, https://doi.org/10.1109/grc.2006.1635905.
[17] T. Zhang, Y. Li, and X. Wang, "Gaussian prior based adaptive synthetic sampling with non-linear sample space for imbalanced learning," *Knowledge-Based Syst.*, vol. 191, p. 105231, 2020, https://doi.org/10.1016/j.knosys.2019.105231.
[18] R. Y. Goh and L. S. Lee, "Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches," *Adv. Oper. Res.*, 2019, https://doi.org/10.1155/2019/1974794.
[19] T. A. Khan, K. A. Kadir, S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 560–569, 2020, https://doi.org/10.14569/IJACSA.2020.0111170.
[20] S. B. Rakhmetulayeva, K. S. Duisebekova, A. M. Mamyrbekov, D. K. Kozhamzharova, G. N. Astaubayeva, and K. Stamkulova, "Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis," *Procedia Comput. Sci.*, vol. 130, pp. 231–238, 2018, https://doi.org/10.1016/j.procs.2018.04.034.
[21] C. O. Truică and C. A. Leordeanu, "Classification of an imbalanced data set using decision tree algorithms," *UPB Sci. Bull. Ser. C Electr. Eng. Comput. Sci.*, vol. 79, no. 4, pp. 69–84, 2017, https://www.scientificbulletin.upb.ro/rev_docs_arhiva/rez57a_274304.pdf.
[22] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A Classification Model for Class Imbalance Dataset Using Genetic Programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019, https://doi.org/10.1109/ACCESS.2019.2915611.
[23] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharm. Res.*, vol. 12, no. 4, pp. 56–66, 2020, https://doi.org/10.31838/ijpr/2020.12.04.013.
[24] R. Kumar and J. Kaur, *Random forest-based sarcastic tweet classification using multiple feature collection*, vol. 163. Springer Singapore, 2020, https://doi.org/10.1007/978-981-13-8759-3_5.
[25] E. Donnellan, S. Aslan, G. M. Fastrich, and K. Murayama, *How Are Curiosity and Interest Different? Naïve Bayes Classification of People's Beliefs*, vol. 34, no. 1. Educational Psychology Review, 2022, https://doi.org/10.1007/s10648-021-09622-9.
[26] Z. Wang, L. Yao, X. Shao, and H. Wang, "A combination of TEXTCNN model and Bayesian classifier for microblog sentiment analysis," *J. Comb. Optim.*, vol. 45, no. 4, pp. 1–22, 2023, https://doi.org/10.1007/s10878-023-01038-1.
[27] T. Sajana and M. R. Narasingarao, "Classification of imbalanced malaria disease using naïve bayesian algorithm,"

Int. J. Eng. Technol., vol. 7, pp. 786–790, 2018, https://doi.org/10.14419/ijet.v7i2.7.10978.
[28] A. M. De Carvalho and R. C. Prati, "Improving kNN classification under Unbalanced Data. A New Geometric Oversampling Approach," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, no. Cmcc, pp. 1–6, 2018, https://doi.org/10.1109/IJCNN.2018.8489411.
[29] G. Tuerhong, M. Wushouer, and D. Zhang, "An Improved K Nearest Neighbor Classifier for High-Dimensional and Mixture Data," *J. Phys. Conf. Ser.*, vol. 1813, no. 1, 2021, https://doi.org/10.1088/1742-6596/1813/1/012026.
[30] P. H. Putra, M. S. Novelan, and M. Rizki, "Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color," *J. Appl. Eng. Technol. Sci.*, vol. 3, no. 2, pp. 126–132, 2022, https://doi.org/10.37385/jaets.v3i2.763.
[31] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta, "A new evidential K-nearest neighbor rule based on contextual discounting with partially supervised learning," *Int. J. Approx. Reason.*, vol. 113, pp. 287–302, 2019, https://doi.org/10.1016/j.ijar.2019.07.009.
[32] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, 2019, https://doi.org/10.1016/j.artmed.2019.101723.
[33] A. S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3853–3864, 2021, https://doi.org/10.1007/s13369-021-05347-7.
[34] H. Lee, J. Kim, and S. Kim, "Gaussian-based SMOTE algorithm for solving skewed class distributions," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 17, no. 4, pp. 229–234, 2017, https://doi.org/10.5391/IJFIS.2017.17.4.229.
[35] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Inf. Sci. (Ny).*, vol. 512, pp. 1214–1233, 2020, https://doi.org/10.1016/j.ins.2019.10.048.
[36] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, https://doi.org/10.1109/TKDE.2008.239.
[37] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," *Procedia Comput. Sci.*, vol. 176, pp. 420–429, 2020, https://doi.org/10.1016/j.procs.2020.08.043.
[38] X. Xie, H. Liu, S. Zeng, L. Lin, and W. Li, "A novel progressively undersampling method based on the density peaks sequence for imbalanced data," *Knowledge-Based Syst.*, vol. 213, p. 106689, 2021, https://doi.org/10.1016/j.knosys.2020.106689.
[39] A. F. Kallappanamatt, "Application of Synthetic Informative Minority Over-Sampling (SIMO) Algorithm Leveraging Support Vector Machine (SVM) On Small Datasets with Class Imbalance," 2018, https://arrow.tudublin.ie/scschcomdis/136/.
[40] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Improving classifiaction performance of fetal umbilical cord using combination of SMOTE method and multiclassifier voting in imbalanced data and small dataset," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 5, pp. 441–454, 2020, https://doi.org/10.22266/ijies2020.1031.39.
[41] W. Badr, "6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)," *Towar. Data Sci.*, 2019, [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

BIOGRAPHY OF AUTHORS

Muhammad Misdram is a lecturer at the Merdeka University Pasuruan and his other position is The Head of the S1 Study Program RPL (Software Engineering) at the faculty of Information Technology. He got MM degree (Master of Management) from College Management IMNI Jakarta. And got M.Kom degree (Master of Computer) from Dian Nuswantoro University Semarang. He is currently taking Doctoral Program S3 Computer Science at Dian Nuswartoro University Semarang and had done special data mining research on little Small imbalaced dataset.

Edi Noersasongko at this time, he is a Rector of the Dian Nuswantoro University Semarang, whereas his last academic position is Professor. History of education were graduated Bachelor of Computer (Local) at Informatic and Computer College, Jakarta-1983., graduated Bachelor of Computers (State) at Informatic and Computer college, Jakarta-1993., graduated Master of Computer at Technology Information Beneratif Indonesia college, Jakarta-1995., graduated Doctor of Economic (S3) Merdeka University Malang-2005., Honorary Doctorate In Educational Information Technology, Teknikal Malaysia Melaka University-2012. Besides, He is a successful businessman, active in various fields of the organization and created many various IT aplication which already got the copyright. Another outstanding success is qualified entepreneurship.

Purwanto is an Assistant Professor in Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia. He received the Ph.D degree from Faculty of Computing and Informatics Multimedia University, Cyberjaya, Malaysia. He has published research papers in reputed international journals and conferences. His current research interests include modelling, time series forecasting, data and text mining, machine learning, soft computing and decision support system.

Muljono holds a Doctor of Electrical Engineering degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia in 2016. He joined an Internship Program at School of Media Science, Tokyo University of Technology Japan in 2014. He received his M.Kom (Informatics) from STTIBI Jakarta, Indonesia in 2001 and He received his B.Sc (Mathematics) from Universitas Diponegoro (UNDIP) in 1996. He is currently an associate professor at Informatics Engineering Department in Dian Nuswantoro University, Semarang, Indonesia. His research includes artificial intelligence, machine learning, data mining, data science and natural language processing. He has published over 90 papers in international journals and conferences. He can be contacted at email: muljono@dsn.dinus.ac.id.