IJSA (2024), Vol. 8, No. 1, 70-82
DOI: https://doi.org/10.29244/ijsa.v8i1p70-82

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Effectiveness of SMOTE-ENN to Reduce Complexity in Classification Model

Ines Riantika[1], Bagus Sartono[2], Khairil Anwar Notodiputro[3]

[1,2,3]Department of Statistics and Data Science, IPB University, Indonesia
corresponding author: Inesriantika@apps.ipb.ac.id

Abstract

A failure to produce classification models with high performance might be caused by the dataset's characteristics, such as the between-class overlapping and the class imbalance. The higher the data complexity, the more complicated it is for the algorithm to find good models. Combining the issues of class imbalance and overlapping would make the problem more challenging. To deal with this problem, this re-search implemented a hybrid class-balancing technique named SMOTE-ENN. In this method, observations are added to the minority class in order to achieve a balance in class frequencies. Once this has been completed, the subsequent step is to remove any observations labeled differently, to reduce the degree of overlapping. This research revealed that SMOTE-ENN succeeds in doing that. We employed a random forest method to evaluate performance of classification. In 28 out of 46 cases we investigated, the new datasets generated by SMOTE-ENN could produce models with higher accuracy and it’s work better when the imbalance ratio value on cases is higher.

Keywords: complexity measures, imbalance class, random forest, smote-enn.

1. INTRODUCTION

The performance of classification algorithm is usually suffered from complexity of dataset such as the problem of imbalance and overlapping data. Hance, improving the dataset characteristics is crucial to enhance the performance of classification. These characteristics can include overlapping classes, linearity of bound decisions, and imbalance ratio in the dataset (Lancho *et al.* 2023).

Ho and Basu (2002) introduced a measurement to assess the dataset characteristics by examining the geometrical distribution of data. They grouped this intrinsic into three measurements including overlapping individual feature values, separability classes, geometry, topology, and density of manifold. These measurements were further developed by (Lorena *et al.* 2019) and grouped them into six types of measurement. One of the most common issues found in many fields is the imbalance of classes. The imbalance class exists when the instances from the minority class are lower than the majority class (López *et al.* 2013). Another impact of this complexity is the value of accuracy that appears be high, but it is not the actual accuracy, also known as the paradox of accuracy. There have been several research related to imbalance class issues. These include text classification (Padurariu dan Breaban 2019), credit scoring (Kang *et al.* 2021), food insecurity (Dharmawan *et al.* 2022), poverty prediction (Santoso *et al.* 2018) and medical diagnosis (Khushi *et al.* 2021). The imbalance class technique can be categorized into two main methods: data level method and algorithm level (Santoso *et al.* 2017).

Another complexity issue is overlapping of classes in the dataset. When two or more classes in the dataset do not have a clear boundary and appear to be mixed in the same place, it creates difficult for the classifier to correctly predict the class of the dataset. Even when the best classifier is used, achieving a decent classification result can be complicated. Combining the issues of imbalanced classes and overlapping in a dataset can make predicting even more challenging. These problems can be overcome by applying SMOTE-ENN to dataset.

The SMOTE-ENN method was chosen over other approaches, including SMOTE, random oversampling, and random undersampling, because of its improved ability to produce synthetic data. Fernández et al. (2018), assert that the synthetic minority oversampling or SMOTE method can produce noise synthetic data. Moreover, random undersampling could lengthen the predictive model's training period and eliminate significance information, whereas duplicate samples might raise the chance of overfitting in the case of random oversampling (Wongvorachan *et al.* 2023).

SMOTE-ENN is a two-step method that combines oversampling and undersampling techniques. First, it generates synthetic observations from the minority class then it removes instances from majority class. This approach excludes the instances if they have the nearest neighborhood from different class. Sasada *et al.* (2020) claimed that the SMOTE-ENN has the benefit of reducing and eliminating the noise from the dataset. The related work about the imbalance dataset and assessing the complexity measures are (Azhar *et al.* 2022). That research shows that SMOTE-ENN gave an excellent performance for classification model with the measurement in this research using two measurements N1 and T1. The measurements were chosen based on the consideration that these two measures can be used on categorical and numerical data types.

In this paper, the effectiveness of the SMOTE-ENN algorithm was measured by N1 and T1 values investigated. Both the values represent the degree of complexity of data in terms of class overlapping in the dataset. We use data before and after the class balancing technique and tested using the t-test to get the conclusion that there is a difference in the application of class balancing techniques in reducing overlapping. Furthermore, the both performance of original dataset and SMOTE-ENN is also assessed by the balance accuracy from random forest classification to show the gap values that indicate as improvement performance.

2. MATERIALS AND METHODS

2.1 Data Complexity Measurements

Data complexity is characteristic of the datasets that would impact the machine learning algorithms to produce decent performance models. This intrinsic effect from the dataset is one of the classification problems. The methods aim to identify the intrinsic from the dataset established by Ho and Basu (2002) creating the measurement used for examining the characteristics data from the geometrical data distribution. According to Ho *et al.* (2006), data complexity can be classified into three categories. The latest, The measurement of data complexity being developed into six types of data complexity by (Lorena *et al.* 2019), which are:

1) Feature-based measures.
2) Linearity measures
3) Neighbourhood measures
4) Network Measures
5) Dimensionality measures
6) Imbalance measures

Identifying the types of data complexity present in a dataset would guide improved performance of the classifier algorithm using the proper dataset.

This research employs two measurements to identify overlapping issues in datasets using neighborhood measures. This type of complexity measure aims to identify the pattern of boundary decision that separates different label classes, with higher values indicating more complex and overlapping datasets. The measures used in this paper are divided into categories.

a. N1 (fraction of borderline points)

The measurement identifies overlaps by creating a minimum spanning tree (MST). This involves precisely determining the nearest points by utilizing a distance matrix. The instances would connect and assemble by line and forming the graph. The connections between two-point instances are called edges (E), whereas each point is considered a vertex (V). The distance on the resulting graph is denoted as e_{ij} (Barella *et al.* 2021).

$$N1 = \frac{\sum_{i=1}^{n} I\left((x_i, x_j) \in MST \wedge y_i \neq y_j\right)}{n} \quad (1)$$

The x_i is the instance point in the dataset that is linked by MST, while the class y_i is determined based on the two points that are connected from different classes. Additionally, the number of instances in the dataset is represented by n.

b. T1 (fraction of hyperspheres covering data)

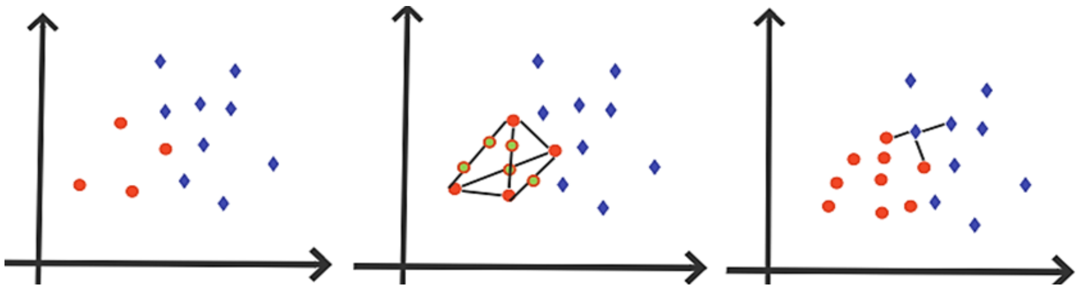
This measurement uses the hyperspheres to identify the instances that have distinct class labels closest. The hyperspheres in every instance have a radius with a zero value and will gradually expand until they touch the hyperspheres from the instance from the other class (Barella *et al.* 2018). According to Barella *et al.* (2021) the instances that belong to the same class and are located nearby are grouped in a single hypersphere.

$$T1 = \frac{\#Hyperspheres(T)}{n} \quad (2)$$

2.2 SMOTE-ENN

The imbalance class is the condition from the dataset that has disproportion class of data represented Douzas *et al.* (2018) .An imbalance class on the dataset will increase the probability that the minority class will fail to predict correctly (Johnson dan Khoshgoftaar 2019). The technique used in this research for handling imbalance class is SMOTE-ENN. This technique is part of the data level method that focuses on balancing the classes by modifying the data distribution by generating new instances (Krawczyk 2016).

SMOTE-ENN combines the SMOTE and ENN and works parallelly. SMOTE is a well-known data-level technique developed to address the issue of overfitting that may arise due to random oversampling of the dataset. By generating new instances of the minority class using an interpolating based on the distance matrix, this technique effectively balances the class distribution. ENN is the undersampling technique that involves balancing the data classes by reducing the majority class using the k-NN rule that is selected based on their proximity to the minority class using a distance matrix where *k* is the number of the closest instances belonging to a different classes(Lu *et al.* 2019). The process from SMOTE-ENN starts with SMOTE, which is used to generate new instances of the minority class. Secondly, ENN is applied to correct and remove the instances that are k nearby and belong to distinct. The capability of SMOTE-ENN to balance distribution data and also tackle the noise may appear after generating the synthetics class from the minority with the interpolation. Puri dan Gupta (2022) said this technique outperforms when the dataset has a higher level imbalance. That is the reason that we are using the SMOTE-ENN in this research. The illustrate of SMOTE ENN is shown on Figure **1**.



a. Imbalance dataset

b. Generating instances using SMOTE

c. Selecting and removing using ENN

Figure 1: Process of SMOTE-ENN

2.3 Data

This research involves 46 public datasets. Among 46 datasets, 14 were obtained from kaggle.com, and 32 were obtained from archive.ics.uci.edu. Those datasets have various levels of imbalance, which are measured by imbalance ratio (IR) values. However, the IR describes a large gap between the majority and minority classes. The higher the value of IR, the higher the imbalance level on the dataset (Zhu *et al.* 2020).

Table 1: List of Datasets

Datasets	Number of instances	Imbalance Ratio	Number of Variables	
Machine Maintenance 3	9697	214.5	6	[1]
Machine Maintenance 2	9747	101.6	6	[1]
Diabetes	7228	44.5	17	[2]

Datasets	Number of instances	Imbalance Ratio	Number of Variables	
Red Wine 1	699	37.8	11	[2]
Company Bankruptcy	6819	30	95	[1]
Thyroid Diseases	3163	19.9	24	[1]
Brain Stroke	4981	19.1	10	[1]
Car Insurance Claim	8789	14.6	42	[1]
Credit	6178	11.6	29	[1]
Personal Loan	5000	9.4	12	[1]
Faults Steel 2	745	9.3	27	[1]
White Wine	1620	8.9	11	[2]
Bank	4119	8.1	20	[1]
Water Quality	7996	7.8	20	[1]
Faults Steel 1	457	7.3	27	[1]
Hepatitis	615	7.2	12	[1]
Mortality	1176	6.4	48	[1]
Stellar	14048	5.5	12	[1]
Insurance	7643	5.1	10	[1]
Glass 3	93	4.5	9	[2]
Telcom Customer Church 1	2323	4.1	30	[1]
Glass 2	87	4.1	9	[2]
Haberman Disease	282	3.7	3	[2]
Immunotherapy	90	3.2	7	[2]
Glass 1	38	3.2	9	[2]
Red Wine 2	837	3.2	11	[2]
Income	9768	2.8	14	[1]
Cirrhosis Disease	418	2.6	18	[1]
Telcom Customer Church 2	6589	2.5	30	[1]
Faults Steel 3	549	2.5	27	[1]
Stroke	14729	2.3	10	[1]
Employee Future	4653	1.9	6	[1]
Pima Indian Diabetes	768	1.9	8	[2]
Ionosphere	351	1.8	31	[2]
Body Signal of Smoking	11138	1.7	25	[1]
Breast Cancer	569	1.7	30	[2]
Car Ownership	500	1.7	7	[1]
Go to Collage	800	1.7	9	[1]
CDR Call Detail	4657	1.6	15	[1]
Machine Maintenance 1	190	1.4	6	[1]
Liver Disorder	345	1.4	5	[2]
Contraception	1473	1.3	9	[2]
Heart Failure	918	1.2	11	[1]
Collage Placement	2966	1.2	7	[1]
Pharyngitis in Children	676	1.2	19	[1]
Heart Disease	303	1.2	13	[1]

Source : [1] Kaggle.com [2] UCI Repository

The series of steps are implemented during the data analysis, as shown in Figure 2: The flow of analysis, and for each dataset in this research produced two values of complexity measures. First, the complexity is calculated from the original datasets. The second value came from the dataset generated by SMOTE-ENN. Those two values are then compared to calculate the impact of SMOTE-ENN in reducing the complexity.

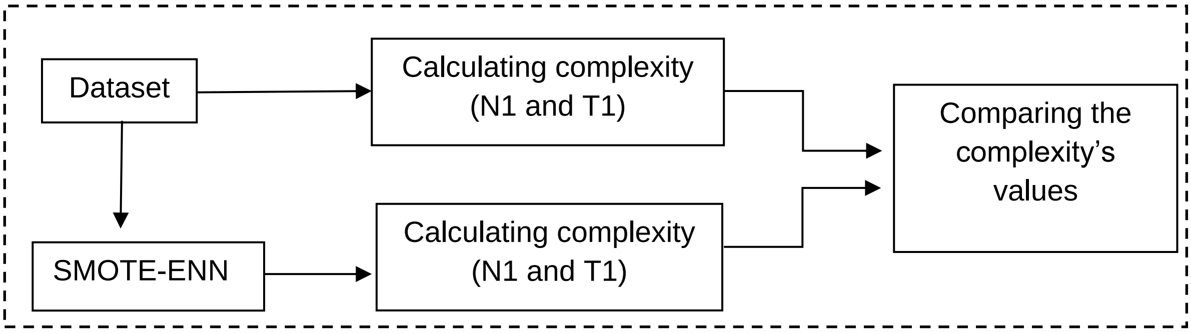

Dataset
SMOTE-ENN
Calculating complexity (N1 and T1)
Calculating complexity (N1 and T1)
Comparing the complexity's values

Figure 2: The flow of analysis

3. RESULT AND DISCUSSION

The application of the SMOTE-ENN to the dataset can lead to a balancing effect by adjusting the number of observations of the class upwards or downwards, as shown in Figure 3. However, the instances generated from this technique may not be equal and still have different proportions of minority and majority classes. In Figure 4, The distribution of status and the changing of complexity T1, we observe that approximately 39.1% of datasets are being reduced, while about 60.9% have increased instances. The reduction in dataset size may indicate that the dataset has possible class overlapping or instances with different classes nearest each other.

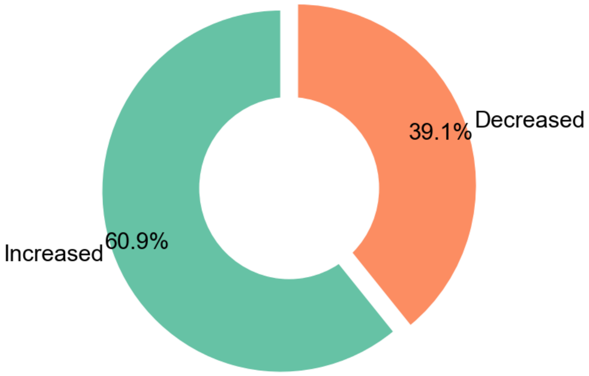


Figure 3: Percentage of the changing number of instances

Figure 4 shows the distribution of datasets on the status of the succession of reduced overlapping after applying the SMOTE-ENN technique. The blue box plot indicates that successful datasets tend to have a significant value of T1 than unsuccessful ones. This information suggests that datasets with a higher value of T1 or more significant complexity value are more likely to be improved through the SMOTE-ENN technique than those with smaller T1 values. The results suggest that the SMOTE-ENN technique effectively reduces overlapping in datasets, particularly for more complex datasets. The result also similar to N1 measurement.

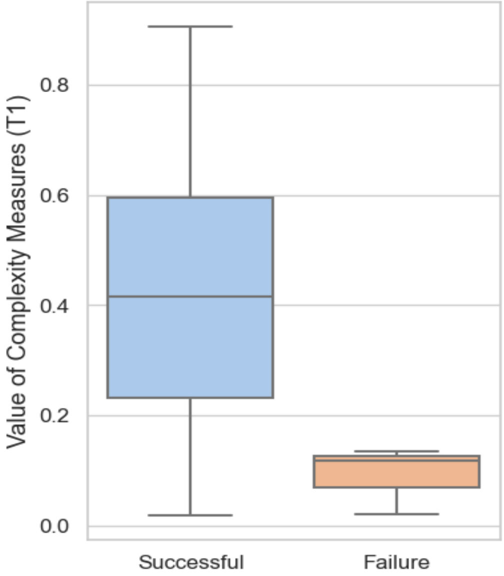
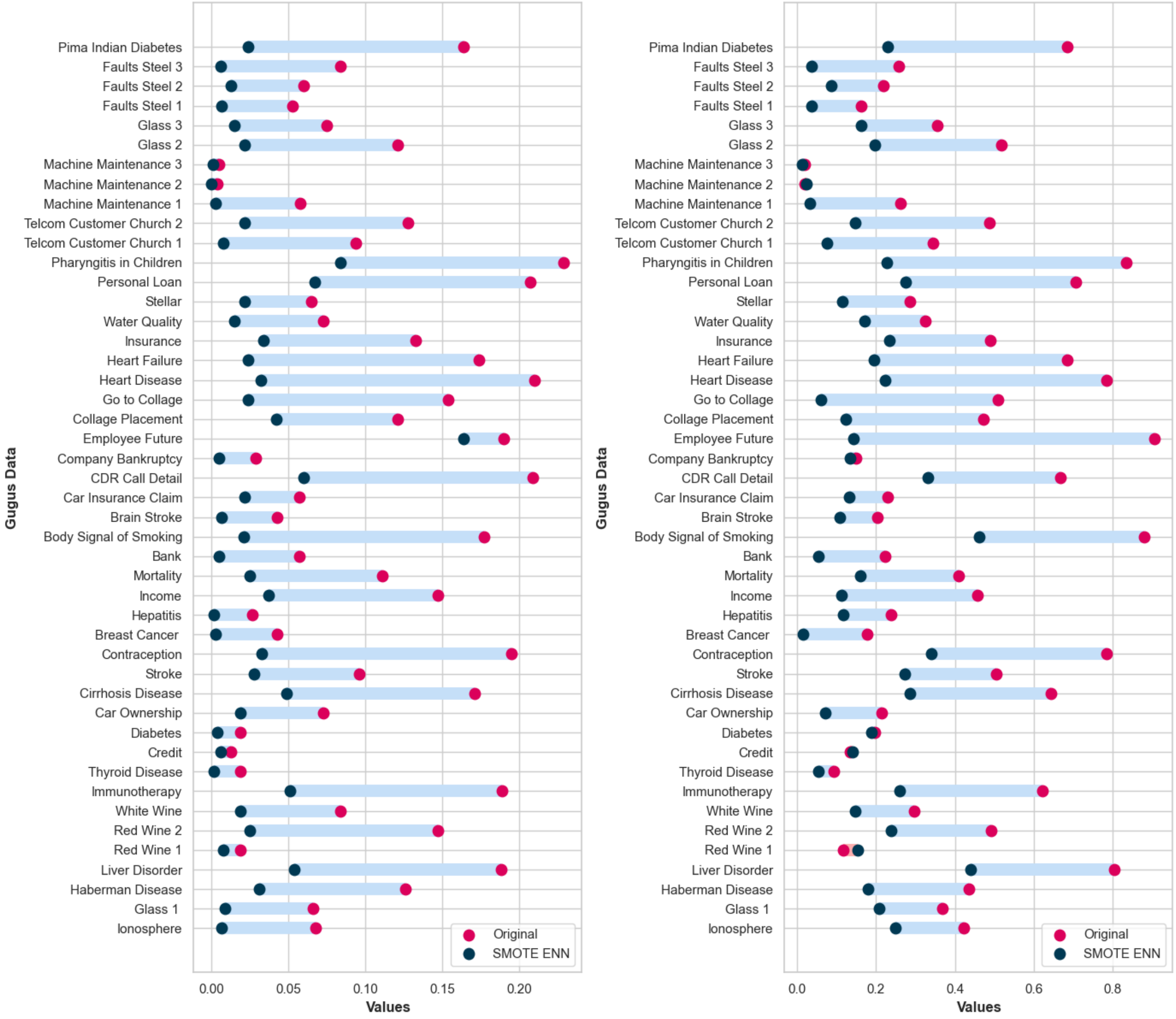


Figure 4: The distribution of status the changing of complexity T1

The effectiveness of the SMOTE-ENN technique is quite remarkable, as evidenced in Figure 5, which is prominent by the blue line on the graph, which means using the SMOTE-ENN could reduce the complexity measure. The analysis reveals that out of the 46 datasets examined, SMOTE-ENN successfully reduced the TI to around 93% of datasets, with only three datasets failing to decrease or 7%. Additionally, the N1 measures demonstrate a consistent trend across all datasets. The results representing SMOTE-ENN are excellent for reducing complexity measures N1 and T1.

The SMOTE-ENN method has been found to be effective in reducing complexity measures (overlapping) on datasets. This is achieved through the utilization of ENN techniques that correct instances that are close to *k* instances with distinct class labels. By addressing the overlapping in the dataset, classifier algorithms could produce improved results and enhance their overall performance.

We apply the t-test to determine whether there is a difference N1 and T1 values in the implementation of SMOTE-ENN in minimizing overlapping data between data collected before and after the technique. This t test can be seen in Table 2, which exhibits the original dataset and after. The difference of the complexity measure from the original dataset is genuinely statistically significant different with the SMOTE-ENN. This could be proved by a p-value less than 0.05, and the different value is positive. This determination proves that the SMOTE-ENN is effective and has successfully reduced the complexity overlapping on the dataset. It is also the same as the past research from Stefanowski (2013), which said the SMOTE-ENN had a good performance when the dataset consists of high overlapping.



a. N1 measurement

b. T1 measurement

Figure 5: The reduction of overlapping measurement

Table 2: Comparing the complexity measures

Complexity	Comparison	t	*p-value*
N1	Original vs SMOTE-ENN	10.76	0.000*
T1	Original vs SMOTE-ENN	9.57	0.000*

To prove the effectiveness of the SMOTE ENN method in reducing complexities (overlapping) on datasets will have the benefit of raising the performance of classification. In this research, we employed random forest classification models to evaluate the performance before (original) and after applying the SMOTE-ENN technique. The results of our experiments on a total of 46 datasets demonstrated that the SMOTE-ENN technique successfully improved the balance accuracy of the classification models.

Table 3 revealed that 26 out of the total datasets showed significant improvement in balance accuracy after applying the SMOTE-ENN technique. Overall, these results suggest that the SMOTE-ENN technique can be a practical approach for addressing overlapping in imbalanced datasets, leading to enhanced classification model performance.

Table 3: Balance accuracy of datasets

Datasets	*Balanced Accuracy* (Original)	*Balanced Accuracy* (SMOTE-ENN)
Machine Maintenance 3	0.500	0.745
Machine Maintenance 2	0.859	0.949
Diabetes	0.509	0.638
Red Wine 1	0.697	0.845
Company Bankruptcy	0.578	0.762
Thyroid Disease	0.841	0.905
Brain Stroke	0.501	0.668
Car Insurance Claim	0.502	0.545
Credit	0.927	0.937
Personal Loan	0.566	0.554
Faults Steel 2	0.948	0.955
White Wine	0.628	0.711
Bank	0.696	0.853
Water Quality	0.829	0.870
Faults Steel 1	0.900	0.927
Hepatitis	0.907	0.940
Mortality	0.563	0.698
Stellar	0.994	0.994
Insurance	0.626	0.783
Glass 3	0.747	0.822
Glass 2	0.593	0.624
Telcom Customer Church 1	0.775	0.853
Haberman Disease	0.553	0.657
Immunotherapy	0.689	0.641
Glass 1	0.921	0.891
Red Wine 2	0.726	0.719
Income	0.766	0.772
Cirrhosis Disease	0.573	0.664
Telcom Customer Church 2	0.814	0.822
Faults Steel 3	0.991	0.932
Stroke	0.973	0.938
Employee Future	0.653	0.650
Pima Indian Diabetes	0.720	0.750
Ionosphere	0.914	0.905
Body Signal of Smoking	0.749	0.753
Breast Cancer	0.954	0.936
Car Ownership	0.915	0.898
Go to Collage	0.896	0.784
CDR Call Detail	0.760	0.678
Machine Maintenance 1	0.946	0.910
Liver Disorder	0.698	0.646
Contraception	0.661	0.689
Heart Failure	0.861	0.824

Datasets	*Balanced Accuracy* (Original)	*Balanced Accuracy* (SMOTE-ENN)
Collage Placement	0.879	0.876
Pharyngitis in Children	0.677	0.626
Heart Disease	0.815	0.725

Table 3**:** Balance accuracy of datasets indicates that the SMOTE-ENN technique fails to improve the balance accuracy when the imbalance ratio is small. However, it effectively improves the balance accuracy when the imbalance ratio is significant. This result suggests that the technique is better suited for datasets with larger imbalance ratios.

The results indicate that the SMOTE-ENN technique successfully reduces the measures for overlapping, namely N1 and T1. However, it is essential to note that the reduction of overlapping measures alone is not sufficient for improving the performance of the random forest classifier using balance accuracy. The imbalance ratio is another crucial factor that needs to be considered. In other words, the effectiveness of the SMOTE-ENN technique in improving balance accuracy is another factor, such as the imbalance ratio of the dataset.

Table 4: Balance accuracy and imbalance ratio

Imbalance Ratio	Balance Accuracy Improved	Balance Accuracy Improved	Total of Datasets
< 2	3	12	15
2 – 10	16	6	22
> 10	9	0	9

For a better understanding of the relationship between the imbalance ratio and balance accuracy performance, we categorize the imbalance ratio level into three parts to demonstrate the severity of the imbalance ratio has been provided in Table 4. It indicates that when the imbalance ratio is less than 2, applying the SMOTE-ENN technique on the random forest classifier only leads to balance accuracy improvement in 3 out of 12 datasets. An imbalance ratio of less than 2 suggests that the minority class accounts for nearly half of the total number of instances of the majority class. Moreover, for imbalance ratios ranging from 2 to 10, roughly 16 out of 22 datasets improved. Finally, when the imbalance ratio exceeds 10, the efficacy of the SMOTE-ENN technique in improving balance accuracy becomes limited. This technique can effectively reduce the overlapping class in the dataset. There needs to be more than reducing the overlapping to increase classification performance. When the dataset has a lower imbalance class in this research, IR less than two SMOTE-ENN does not work correctly, and the performance exhibits many datasets have decreasing balance accuracy values. This finding conforms to the expectations of academic discourse, where a thorough examination of the factors that contribute to the performance of a technique is deemed crucial.

4. CONCLUSION AND SUGGESTIONS

The SMOTE-ENN technique effectively addresses overlapping datasets while dealing with class imbalance. Using two measures, N1 and T1, has resulted in good outcomes, where around 43 datasets have shown a reduction in T1 values, and all datasets have seen a reduction in N1 measures. Using paired t-test proves that complexity measures from original datasets are statistically significantly different from the SMOTE-ENN dataset. The result also describe the SMOTE-ENN is effective in reducing the overlapping problem (complexity measurements T1 and N1).

The impact of an imbalance ratio on the performance of random forest is significant where an imbalance ratio has a value of less than two or when the minority class accounts for approximately 50% of the majority class. Many dataset fields require improvement to achieve balance accuracy. In conclusion, the SMOTE-ENN technique works when dealing with a significant imbalance ratio and high levels of overlapping.

The suggestion to this research is that the complexity measurement in this paper only utilized one complexity measure type: neighbourhood measures with the name measurement T1 and N1. These types of complexity measures imply describing the overlapping class in the dataset. To get a better conclusion and extensive information, the complexity of overlapping is necessary to observe with the other complexity measurements simultaneously.

This research implies that there is a need to evaluate other class-balancing techniques that have the effect of simultaneously reducing various types of data complexity measures and improving classification model performance.

REFERENCES

Azhar NA, Mohd Pozi MS, Mohamed Din A, Jatowt A. 2022. An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis. *IEEE Trans Knowl Data Eng*. PP c:1. doi:10.1109/TKDE.2022.3179381.

Barella VH, Garcia LPF, de Souto MCP, Lorena AC, de Carvalho ACPLF. 2021. Assessing the data complexity of imbalanced datasets. *Inf Sci (Ny)*. 553:83–109. doi:10.1016/j.ins.2020.12.006.

Barella VH, Garcia LPF, De Souto MP, Lorena AC, De Carvalho A. 2018. Data Complexity Measures for Imbalanced Classification Tasks. *Proc Int Jt Conf Neural Networks*. 2018-July. doi:10.1109/IJCNN.2018.8489661.

Dharmawan H, Sartono B, Kurnia A, Hadi AF, Ramadhani E. 2022. a Study of Machine Learning Algorithms To Measure the Feature Importance in Class-Imbalance Data of Food Insecurity Cases in Indonesia. *Commun Math Biol Neurosci*. 2022:1–25. doi:10.28919/cmbn/7636.

Douzas G, Bacao F, Last F. 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf Sci (Ny)*. 465:1–20. doi:10.1016/j.ins.2018.06.056.

Ho TK, Basu M. 2002. Complexity measures of supervised classification problems. *IEEE Trans Pattern Anal Mach Intell*. 24(3):289–300. doi:10.1109/34.990132.

Ho TK, Basu M, Law MHC. 2006. Measures of Geometrical Complexity in Classification Problems. *Data Complex Pattern Recognit*.(3):1–23. doi:10.1007/978-1-84628-172-3_1.

Johnson JM, Khoshgoftaar TM. 2019. Survey on deep learning with class imbalance. *J Big Data*. 6(1). doi:10.1186/s40537-019-0192-5.

Kang Y, Jia N, Cui R, Deng J. 2021. A graph-based semi-supervised reject inference framework considering imbalanced data distribution for consumer credit scoring. *Appl Soft Comput*. 105:107259. doi:10.1016/j.asoc.2021.107259.

Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, Yang X, Reyes MC. 2021. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*. 9:109960–109975. doi:10.1109/ACCESS.2021.3102399.

Krawczyk B. 2016. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell*. 5(4):221–232. doi:10.1007/s13748-016-0094-0.

Lancho C, Martín De Diego I, Cuesta M, Aceña V, Moguerza JM. 2023. Hostility measure for multi-level study of data complexity. *Appl Intell*. 53(7):8073–8096. doi:10.1007/s10489-022-03793-w.

López V, Fernández A, García S, Palade V, Herrera F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf Sci (Ny)*. 250:113–141. doi:10.1016/j.ins.2013.07.007.

Lorena AC, Garcia LPF, Lehmann J, Souto MCP, Ho TKAM. 2019. How complex is your classification problem?: A survey on measuring classification complexity. *ACM Comput Surv*. 52(5). doi:10.1145/3347711.

Lu T, Huang Y, Zhao W, Zhang J. 2019. The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN. *Proc IEEE 7th Int Conf Comput Sci Netw Technol ICCSNT 2019*., siap terbit.

Padurariu C, Breaban ME. 2019. Dealing with data imbalance in text classification. *Procedia Comput Sci*. 159:736–745. doi:10.1016/j.procs.2019.09.229.

Puri A, Gupta MK. 2022. Improved Hybrid Bag-Boost Ensemble with K-Means-SMOTE-ENN Technique for Handling Noisy Class Imbalanced Data. *Comput J*. 65(1):124–138. doi:10.1093/comjnl/bxab039.

Santoso B, Wijayanto H, Notodiputro KA, Sartono B. 2017. Synthetic over Sampling Methods for Handling Class Imbalanced Problems : A Review. Di dalam: *IOP Conference Series: Earth and Environmental Science*. Volume ke-58. Institute of Physics Publishing.

Santoso B, Wijayanto H, Notodiputro KA, Sartono B. 2018. A Comparative Study of Synthetic Over-sampling Method to Improve the Classification of Poor Households in Yogyakarta Province. *IOP Conf Ser Earth Environ Sci*. 187(1):0–18. doi:10.1088/1755-1315/187/1/012048.

Sasada T, Liu Z, Baba T, Hatano K, Kimura Y. 2020. A resampling method for imbalanced datasets considering noise and overlap. *Procedia Comput Sci*. 176:420–429. doi:10.1016/j.procs.2020.08.043.

Stefanowski J. 2013. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. *Smart Innov Syst Technol*. 13:277–306. doi:10.1007/978-3-642-28699-5_11.

Wongvorachan T, He S, Bulut O. 2023. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Inf*. 14(1). doi:10.3390/info14010054.

Zhu R, Guo Y, Xue JH. 2020. Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognit Lett*. 133:217–223. doi:10.1016/j.patrec.2020.03.004.