Handling Imbalance Data Using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification

Muhammad Abdul Latief, Luthfi Rakan Nabila, Wildan Miftakhurrahman, Saihun Ma'rufatullah, Henri Tantyoko
Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia

Article Info

Article history:

Received January 07, 2024
Revised January 16, 2024
Accepted February 04, 2024

Keywords:

Hybrid Sampling
Lung Cancer
Imbalance Data
Resampling
Random Forest

ABSTRACT

The classification problem is a problem typically handled or resolved using machine learning. When there is an imbalance in the classes within the data, machine learning models tend to overclassify a greater number of classes. Due to the issue, the model will have low accuracy in a few classes and high accuracy in many classes. Most of the data has the same number of classes, but if the difference is too great, it will differ. The issue of data imbalance is also evident in the data on lung cancer, where there are 283 positive classes and 38 negative classes. Therefore, **this research aimed** to use a hybrid sampling technique, combining Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN) and Random Forest, to balance the data of lung cancer patients who experience class imbalance. **This research method involved the SMOTE-ENN** preprocessing method to balance the data. The Random Forest method was used as a classification method to predict lung cancer by dividing training data and testing 10-fold cross-validation. **The results of this study** showed that using SMOTE-ENN with Random Forest has the best performance compared to SMOTE and without oversampling on all metrics used. **The conclusion** was using the SMOTE-ENN hybrid sampling technique with the Random Forest model significantly improves the model's ability to identify and classify data.

Corresponding Author:

Muhammad Abdul Latief,
Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia.
Email: 21110002@ittelkom-pwt.ac.id

How to Cite: M. Latief, L. Nabila, W. Miftakhurrahman, S. Marufatullah, and H. Tantyoko, "Handling Imbalance Data using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 3, no. 1, pp. 11-18, March. 2024. doi: 10.30812/ijecsa.v3i1.3758.

1. INTRODUCTION

Artificial intelligence is one of the implementations of the rapid development of technology [1]. Artificial intelligence improves the performance of computers/software to obtain and process information by adopting and imitating human intelligence. One of the artificial intelligence applications is machine learning, which focuses on developing systems capable of self-learning without the need to reprogram continuously [2]. Machine learning is computer programming that aims to achieve certain criteria by utilizing training data or experience [3]. One example of a problem usually handled or solved by machine learning is the classification problem. Classification is a machine learning model that predicts the appropriate category or label [4]. Classification is used to estimate a class in data that is unknown beforehand [5]. Until now, many algorithms have been developed for classification, but some problems often become obstacles in classification, namely unbalanced classes in the data [6]. Machine learning models tend to over-classify a larger number of classes when there is a class imbalance in the data [7]. The impact of the problem is that the model will have high accuracy on a large number of classes and low accuracy on a small number of classes [8, 9]. The reality is that most of the data has the same number of classes, but it will be different if the difference is too large. The problem of data imbalance also appears in the lung cancer data, with the number of positive classes 283 and negative classes 38. Therefore, researchers focus on the imbalance problem that exists in the data.

Various techniques have been proposed to address the issue of imbalanced data, with resampling being one of the prominent approaches [10]. Resampling methods aim to rebalance the class distribution by manipulating the dataset through different sampling algorithms, thus enabling better training of classification models. These techniques generally fall into three categories: undersampling, oversampling, and hybrid sampling. For instance, a previous study [11] utilized undersampling in combination with the Random Forest K-Fold algorithm to mitigate class imbalance, resulting in improved performance metrics such as AUC scores exceeding 0.5. Similarly, another investigation [12] compared the effectiveness of oversampling techniques, particularly SMOTE, with traditional methods. The study reported significant enhancements across various evaluation metrics, including accuracy, sensitivity, precision, G-Mean, F1-score, specificity, and Youden's Index. However, despite these advancements, neither undersampling nor oversampling alone can fully address the complexities of imbalanced datasets. In contrast, our proposed approach combines the strengths of both undersampling and oversampling through the SMOTE-ENN method, offering a more comprehensive solution for lung cancer classification. By synthetically generating minority class instances while simultaneously removing noisy samples, SMOTE-ENN enhances the discriminative power of the model, resulting in improved performance and robustness against class imbalance. Therefore, our study aims to investigate the efficacy of this combined approach in handling imbalanced data and its impact on lung cancer classification performance.

In this study, researchers tried to use hybrid sampling techniques to handle class imbalances in lung cancer patient data. This study tests the SMOTE-ENN algorithm and performs a combination with Random Forest in balancing classes according to suggestions in research [13]. Research [13] combined SMOTE and Random Forest for imbalanced data. As a result, the combination of random forest and SMOTE improved by 5% accuracy and 39% sensitivity compared to random forest without SMOTE. The use of the SMOTE-ENN hybrid sampling method with Random Forest in the classification of lung diseases **has not been carried out by previous research**. So, **the difference between this research** and research [13] lies in the resampling technique used to handle data imbalance. Research [13] uses oversampling techniques, while this research uses hybrid sampling techniques that combine undersampling and oversampling. The data used in this study is lung cancer patient data obtained from the data provider website, Kaggle. The data has a problem with class imbalance with a class ratio of 283 and 38. From the problem of imbalance, researchers use hybrid sampling techniques to perform resampling, in this case, SMOTE-ENN, which is then modeled with the Random Forest classification algorithm. Therefore, **this research aims** to utilize hybrid sampling techniques to handle imbalanced lung cancer and improve the accuracy of random forest classification. **This study contributes** to using hybrid sampling techniques (SMOTE-ENN) to handle class imbalance in lung cancer patient data. The combination of undersampling and oversampling is expected to provide a more comprehensive solution than using either technique separately, resulting in increased performance and robustness in lung cancer classification with the Random Forest algorithm.

2. RESEARCH METHOD

The effectiveness of a cancer prediction system can help people to know their cancer risk at a lower cost and can also help people to make the right decision based on their cancer risk status. This research has a process flow chart framework to achieve these goals, as in Figure 1.

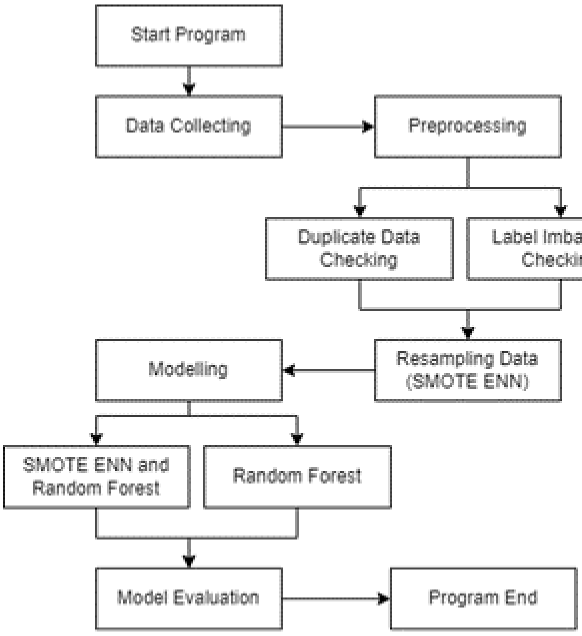

Figure 1. Flow of Research Method

2.1. Data Collecting

The quality and accuracy of the data used often determine the success of a study. Therefore, the first crucial step in the success of the research is data collection. In this context, the researcher has collected the diabetes dataset, a step that requires care and caution. The data source used comes from Kaggle, a leading platform that provides a variety of datasets for analysis and research purposes. The selection of this dataset was not done haphazardly but through careful consideration to ensure that the data obtained was in line with the research objectives and had a reliable quality. Data retrieval is a mechanical downloading process and involves a deep understanding of the dataset's characteristics. Researchers need to ensure that the data retrieved has a high relevance to the research focus and identify potential biases or anomalies that may appear in the dataset. This step sets a solid foundation for the rest of the research journey, ensuring a strong foundation before entering further analysis.

2.2. Preprocessing

In the preprocessing stage, lung cancer data is processed from its raw form into a format that is ready to be trained by classification models to avoid potential problems that could interfere with classification results [10]. Steps involve checking and removing duplicate data, evaluating label imbalance, and using hybrid resampling techniques, specifically SMOTE ENN, to align the number of instances between the majority and minority classes [11]. The result of this process is a balanced dataset, ensuring that the model to be trained can learn well from both classes and produce accurate classification results.

2.3. Modeling

The modeling stage is the third stage after the data collection and preprocessing. In this data, clean data will be modeled using the Random Forest algorithm using the scikit-learn library in Python. Model selection is based on data characteristics and analysis objectives, considering model performance and interpretability. However, before modeling, the data will be divided into training and test data using the 10-fold cross-validation technique [12].

2.4. Model Evaluation

The last stage of this research is to evaluate the model's performance by measuring the accuracy of the Random Forest algorithm in classifying lung cancer. The evaluation stage uses three measurement metrics: accuracy, recall or sensitivity, and specificity. The accuracy value describes how accurately the system can classify the accuracy results. It describes how accurately the system can classify the data correctly. In other words, the accuracy value is the ratio of correctly classified data to the overall data. The three techniques use different approaches, as shown in Equations (1), (2), and (3). The accuracy value can be obtained with equation (1). The accuracy result describes how accurately the system can classify. Then, the specificity value describes the number of correctly classified positive category data divided by the total data categorized as positive. Specificity is obtained in equation (3). Meanwhile, the sensitivity value shows how much of the data from the positive category is correctly classified by the system. The sensitivity value is obtained by Equation (2).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

Where TP is True Positive (True detected positive data), TN is True Negative (Number of correctly detected negative data), FP is False Positive (Negative data but detected as positive data), and FN is False Negative (Positive data but detected as negative data).

3. RESULT AND ANALYSIS

3.1. Data Collecting

In the initial stage of the data collection process, secondary data in the form of lung cancer data from the Kaggle site was obtained. The lung cancer dataset obtained from Kaggle has 309 data and 15 attributes. The attributes of lung cancer in the dataset can be seen in Table 1.

Table 1. Attributes of the Lung Cancer Disease Dataset

No.	Attributes	Description
1.	Gender	Gender is an attribute of the patient's gender
2.	Age	Age is the patient's age attribute
3.	Smoking	Attributes that describe whether the patient is a smoker
4.	Yellow_Fingers	Attributes in the form of a question whether the patient has yellow fingers
5.	Anxiety	Excessive panic when breathing out of breath rhythm
6.	Peer_Pressure	Psychological stress or feeling pressured by the environment (shortness of breath in crowds)
7.	Chronic Disease	Having a chronic disease
8.	Fatigue	Rapid atigue during daily activities
9.	Allergy	Allergic disease
10.	Wheezing	Breathing sounds
11.	Alcohol Consuming	History of alcohol consumption
12	Coughing	Coughing is an attribute of coughing
13.	Shortness of Breath	Shortness of breath is an attribute of shortness of breath
14.	Swallowing Difficulty	Difficulty swallowing is an attribute of superficial difficulty
15.	Lung Cancer	Prediction Class

3.2. Preprocessing

In this stage, a data check is conducted to find whether there is data that has the same value. After eliminating duplicate data, a check is made to see if there is an imbalance of labels in the data. The results show that the majority class in the lung cancer data has 238 instances, while the minority class has only 38 instances. To solve this imbalance problem, a resampling process is performed on the data. Resampling in this study uses a hybrid technique, which combines oversampling and undersampling. The hybrid method implemented is SMOTE ENN, which aims to align the number of examples between the majority class and the minority class so that both have a balanced or not too different distribution, as in Table 2.

Table 2. Comparison of Number of Classes of Resampling Methods

Methods	Number of Instances	
	Positive Class	Negative Class
Without SMOTE-ENN	238	38
SMOTE (Previous Research)	238	238
SMOTE-ENN	214	173

3.3. Modeling

After the data is processed and cleaned, the next step is modeling using Random Forest. However, before modeling, the data is first divided using K-Fold Cross-validation with the number K as 10. In the modeling process, there are two modeling schemes; the first is Random Forest without imbalance handling with SMOTE-ENN and modeling using SMOTE-ENN + Random Forest. The 2 model schemes will be compared in the evaluation process to determine the best model.

3.4. Model Evaluation

In this process, the models made, namely the Random Forest model without SMOTE-ENN and Random Forest with SMOTE-ENN, are evaluated using accuracy score, sensitivity, and specificity. The accuracy scores of both models can be seen in Table 3.

Table 3. Metric Score Model

Model	Accuracy	Recall (Sensitivity)	Specificity
Random Forest	90.2%	90.2%	89.7%
SMOTE + Random Forest (Previous Research)	94.1%	94.5%	93.7%
SMOTE-ENN + Random Forest	99.7%	99.7%	99.7%

Table 3 shows that Random Forest has an accuracy of 90.2%, recall (sensitivity) of 90.2%, and specificity of 89.7%. This shows that the Random Forest model has good performance in data classification. However, when looking at the model using SMOTE-ENN + Random Forest, all metrics show a noticeable improvement. The accuracy reached 99.7%, as well as the recall and specificity. Table 3 also shows the considerable difference between SMOTE+RF and SMOTE-ENN+RF, with a 5.6% difference in accuracy. **This study found** that using the SMOTE-ENN hybrid sampling technique with the Random Forest model significantly improves the model's ability to identify and classify data.

In Figure 2, the Random Forest method without SMOTE correctly predicted the cancer grade in 22 cases out of 38 data. In comparison, the non-cancer category was correctly predicted in 227 cases out of 238 data. In Figure 3, the Random Forest method with SMOTE-ENN correctly predicted the cancer grade in 214 out of 238 cases, while the non-cancer grade was 172 out of 238 cases. So, the use of SMOTE-ENN can help improve the efficiency of the Random Forest accuracy classification method, **this is in line with the research results** [14–16].

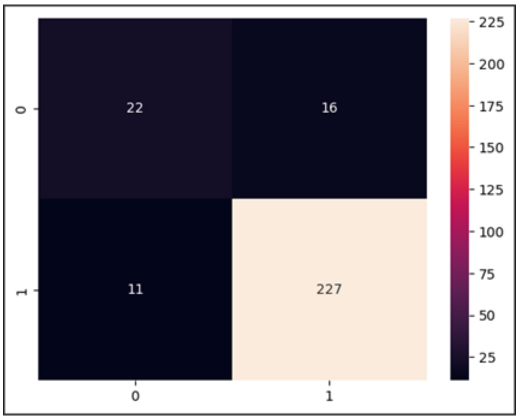

Figure 2. Confusion Matrix of Random Forest

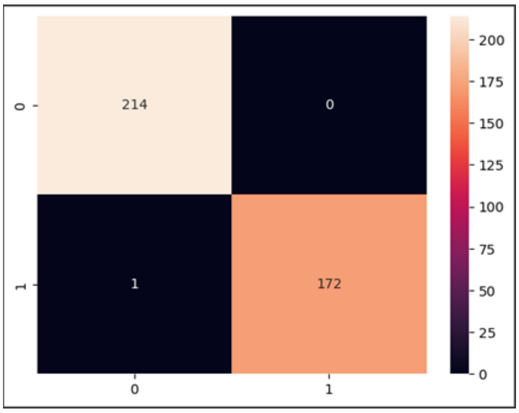

Figure 3. Confusion Matrix of SMOTE-ENN with Random Forest

4. CONCLUSION

After conducting the testing process, it was found that combining the SMOTE-ENN method with Random Forest provides accuracy, sensitivity, and specificity results of 99.7%. Meanwhile, without SMOTE-ENN, the accuracy, sensitivity, and specificity only reached 90.2%, 90.2%, and 89.7%, respectively. The combination of the SMOTE-ENN method with Random Forest improved the model's performance with an increase in accuracy of 9.5% and sensitivity of 9.5% when compared to without the use of SMOTE-ENN. In addition, SMOTE-ENN+RF also has better accuracy than the previous study that used SMOTE+RF, with an accuracy difference of 5.6%. This confirms that the application of SMOTE-ENN can significantly improve the predictive ability of lung cancer through the Random Forest method.

5. DECLARATIONS

AUTHOR CONTIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

-

COMPETING INTEREST

The authors declare no conflict of interest in this article.

REFERENCES

[1] K. R. Ririh, N. Laili, A. Wicaksono, and S. Tsurayya, "Studi Komparasi dan Analisis SWOT pada Implementasi Kecerdasan Buatan (Artificial Intelligence) di Indonesia," *J@ti Undip: Jurnal Teknik Industri*, vol. 15, no. 2, pp. 122–133, Jun. 2020, publisher: Departemen Teknik Industri, Fakultas Teknik, Universitas Diponegoro. [Online]. Available: https://ejournal.undip.ac.id/index.php/jgti/article/view/29183

[2] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine," *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, vol. 12, no. 1, pp. 67–80, May 2020. [Online]. Available: http://ojs.stmik-im.ac.id/index.php/INFORMASI/article/view/48

[3] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, Nov. 2020. [Online]. Available: https://journal.stekom.ac.id/index.php/E-Bisnis/article/view/247

[4] N. Salim, "Penggunaan Jaringan Syaraf Tiruan Untuk Optimasi Kontruksi Bendung Tyrol Plat Berlubang (Study Kasus Pemodelan Bendung Tyrol Plat Berlubang, Provinsi Ankara, Turkey)," *JUSTINDO (Jurnal Sistem dan Teknologi Informasi Indonesia)*, vol. 7, no. 1, pp. 50–58, Mar. 2022. [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/5898

[5] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes,” *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021. [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/2937

[6] D. Pramadhana, R. Rendi, and R. Robiyanto, “Peningkatan Algoritma J48 Untuk Klasifikasi Hasil Prestasi Mahasiswa Selama Proses Pembelajaran Secara Daring Menggunakan CFS Dan Adaboost,” *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, vol. 5, no. 1, pp. 17–26, Dec. 2022. [Online]. Available: http://journal.ittelkom-pwt.ac.id/index.php/inista/article/view/853

[7] T. Pan, J. Zhao, W. Wu, and J. Yang, “Learning imbalanced datasets based on SMOTE and Gaussian distribution,” *Information Sciences*, vol. 512, pp. 1214–1233, Feb. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0020025519310187

[8] D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” *Information Sciences*, vol. 505, pp. 32–64, Dec. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0020025519306838

[9] E. F. Swana, W. Doorsamy, and P. Bokoro, “Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset,” *Sensors*, vol. 22, no. 9, p. 3246, Apr. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/9/3246

[10] M. P. Pangestika, I. M. Sumertajaya, and A. Rizki, “Penerapan Synthetic Minority Oversampling Technique pada Pemodelan Regresi Logistik Biner terhadap Keberhasilan Studi Mahasiswa Program Magister IPB,” *Xplore: Journal of Statistics*, vol. 10, no. 2, pp. 152–166, May 2021. [Online]. Available: https://stat.ipb.ac.id/journals/index.php/xplore/article/view/238

[11] R. D. Fitriani, H. Yasin, and T. Tarno, “Penanganan Klasifikasi Kelas Data Tidak Seimbang dengan Random Oversampling pada Naive Bayes (Studi Kasus: Status Peserta KB IUD I Kabupaten Kendal),” *Jurnal Gaussian*, vol. 10, no. 1, pp. 11–20, Feb. 2021, number: 1 Publisher: Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro. [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/article/view/30243

[12] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier Berdasarkan Data Tidak Seimbang,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022. [Online]. Available: https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/1726

[13] C. Michael Lauw, H. Hairani, I. Saifuddin, J. Ximenes Guterres, M. Maariful Huda, and M. Mayadi, “Combination of Smote and Random Forest Methods for Lung Cancer Classification,” *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 2, no. 2, pp. 59–64, Sep. 2023. [Online]. Available: https://journal.universitasbumigora.ac.id/index.php/IJECSA/article/view/3333

[14] I. Yulianti, A. Rahmawati, and T. Mardiana, “The Effectiveness Analysis of Random Forest Algorithms with Smote Technique in Predicting Lung Cancer Risk,” *Jurnal Riset Informatika*, vol. 4, no. 2, pp. 207–214, Mar. 2022. [Online]. Available: https://ejournal.kresnamediapublisher.com/index.php/jri/article/view/385

[15] H. Hairani and D. Priyanto, “A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. [Online]. Available: http://thesai.org/Publications/ViewPaper?Volume=14&Issue=8&Code=IJACSA&SerialNo=64

[16] H. Hairani, A. Anggrawan, and D. Priyanto, “Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link,” *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 1, p. 258, Feb. 2023. [Online]. Available: http://joiv.org/index.php/joiv/article/view/1069

[This page intentionally left blank.]