Combination of Smote and Random Forest Methods for Lung Cancer Classification

Christopher Michael Lauw[1], Hairani Hairani[1], Ilham Saifudin[2], Juvinal Ximenes Guterres[3], Muhammad Maariful Huda[4], Mayadi Mayadi[5]

[1]Universitas Bumigora, Mataram, Indonesia
[2]Universitas Muhammadiyah Jember, Jember, Indonesia
[3]Universidade Oriental Timur Lorosae, Unital Becora Dili, Timor Leste
[4]Politeknik Angkatan Darat, Malang, Indonesia
[5]University Teknologi Mara, Selangor, Malaysia

Article Info	ABSTRACT
Article history: Received August 30, 2023 Revised September 08, 2023 Accepted September 20, 2023 **Keywords:** Data Mining Lung Cancer Prediction Method Random Forest	Lung cancer is a network of cells that grow abnormally in the lungs. Lung cancer has four severity levels, namely stages 1 to 4. If lung cancer is not treated quickly, it is at risk of causing death. This research aimed to combine Synthetic Minority Over-sampling (Smote) and Random Forest methods for lung cancer classification. The method used was a combination of Smote and Random Forest. Smote was used to balance the data, while Random Forest was used to classify lung cancer data. The results showed that the combination of Smote and Random Forest methods obtained an accuracy of 94.1%, sensitivity of 94.5, and specificity of 93.7%. Meanwhile, without Smote, the accuracy is 89.1%, sensitivity is 55%, and specificity is 94.5%. The use of Smote can improve the performance of the Random Forest classification method based on accuracy and sensitivity. There was an increase of 5% in accuracy and a 39% increase in sensitivity.

Corresponding Author:

Hairani Hairani,
Universitas Bumigora, Mataram, Indonesia.
Email: hairani@universitasbumigora.ac.id

How to Cite: 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. 63 - 70, Sep. 2023. doi: 10.30812/ijecsa.v2i2.3333.

1.   INTRODUCTION

Lung cancer is an abnormal cell tissue that grows in the lungs. The problem found is that many patients are not saved due to lung cancer because it is not detected early [1]. The cause of the slow detection of lung cancer is the risk of causing death [2]. Early detection of lung cancer must be done to get fast and proper treatment to reduce the risk of more severe [3]. Many lung cancer screening applications can be utilized in today's technology era. Early detection of lung cancer independently can be done by entering the symptoms experienced in the application installed on the smartphone. The problem is that the level of accuracy of the resulting diagnosis does not have high accuracy. Therefore, this research not only focuses on developing an early screening application for lung cancer but also considers a very high level of accuracy. To get high-accuracy results, looking at the complexity of the lung disease dataset used is necessary. This research uses a lung cancer dataset obtained from Kaggle with a positive number of cancer of 238 instances and negative cancer of 38 instances. The lung cancer dataset has an unbalanced data problem, with more positive than negative cancer classes. This will cause the prediction method to recognize more cancer-positive classes than cancer-negative classes so that the performance of the prediction method is low.

Cancer prediction has been widely used as an object of research by previous researchers with various prediction methods, such as research [4] using the Naive Bayes method for lung cancer classification with an accuracy of 73%. Research [5] using the Random Forest method with principal component analysis (PCA) feature selection obtains an accuracy of 90.1%. Research [6] uses forward selection to improve the performance of the K-NN method on cancer prediction by obtaining an accuracy of 91.43%. Research [7] uses the Stochastic Gradient Boosting (SGD) method for cancer prediction with an accuracy of 63%. Research [8] uses an ensemble learning classification approach with several classification methods for lung cancer prediction. Some classification methods used are Multilayer Perceptron (MLP), Support Vector Machine (SVM), Logistic Regression, Decision Tree, KNN, and Random Forest. The Ensemble classification method obtained an accuracy of 85%, recall of 89%, and precision of 85%, while the Random Forest method obtained an accuracy of 88%, precision of 88%, and recall of 91%. The study results show that the Random Forest method performs better than the Ensemble method based on accuracy, recall, and precision. Research [9] uses Smote to improve the performance of the Extreme Learning method in predicting postoperative life expectancy for breast cancer. His research shows that combining Smote and Extreme Learning methods can get an accuracy of 85.22% and an F-measure of 86.4%. Research [10] uses Forward Chaining and Certainty Factor methods to diagnose rheumatic diseases with an accuracy of 80%. Research [11] used the Smote-Tomek Link oversampling method to improve the performance of the Random Forest method in predicting diabetes with an accuracy of 86.4%, sensitivity of 88.2%, and F1-measure of 85.1%.

Based on several previous studies that have used various approaches for predicting lung cancer, several gaps can be concluded from previous research, such as the level of accuracy is not optimal and solving the problem of unbalanced data on lung disease prediction has not been done. Therefore, it is necessary to improve the performance of classification methods in lung cancer prediction by solving the problem of unbalanced data on lung cancer disease data. An approach that can be used to solve unbalanced data on lung cancer data is Smote [12], 13. The SMOTE method is used to overcome data imbalance by generating artificial data in the minority class so that the number is the same as the majority class [14]. At the same time, the classification method used is Random Forest because it performs better than several other classification methods in lung cancer prediction [8]. Therefore, this research aims to combine Smote and Random Forest methods for lung disease prediction to obtain high accuracy.

2.   RESEARCH METHOD

The research framework is shown in Figure 1. In the initial stage, the data collection process obtains lung cancer data from the Kaggle site. The lung cancer dataset obtained from Kaggle has 276 data and 15 attributes. Lung cancer attributes can be seen in Table 1.
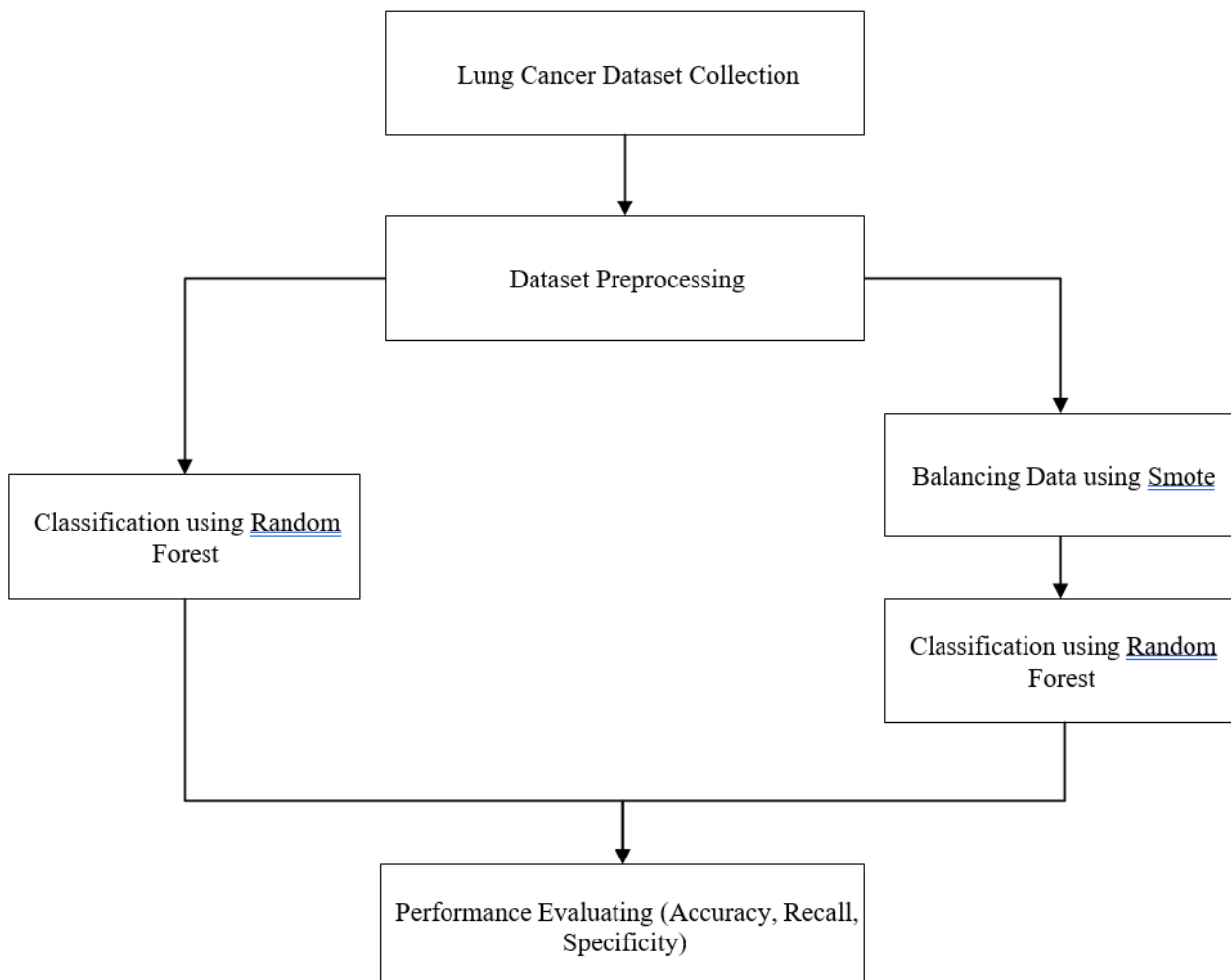
Figure 1. Research Framework

Table 1. Attributes of the Lung Cancer Dataset

No	Attributes	Description
1	Gender	Gender is an attribute of the patient's sex
2	Age	Age is the age attribute of the patient
3	Smoking	An attribute that describes 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	Tired quickly 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 coughs
13	Shortness Of Breath	Shortness of Breath is an attribute of shortness of breath
14	Swallowing Difficulty	Swallowing difficulty is an attribute of shallowing difficulty
15	Lung Cancer	Prediction Class

The next stage is data processing, which is needed to transform raw data into better-quality data so that the performance of the

classification method is better. The processing part uses sampling techniques to balance the data on the lung cancer dataset because the majority class has 238 instances, and the minority class has 38 instances. The Smote method balances the minority class so that the number equals the majority class. The Smote method generates new synthetic data based on the number of nearest neighbors between minority data. The third stage is implementing the Random Forest classification method for predicting lung cancer data by dividing training and testing data based on 10-fold cross-validation. The last stage is a performance evaluation that is needed to determine the level of accuracy of the Random Forest method in predicting lung cancer. Performance testing based on accuracy, recall, and specificity using Equation (1), (2),dan (3).

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

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

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

3.   RESULT AND ANALYSIS

This section will present the research results obtained at each stage. The research stages start from collecting lung cancer data, preprocessing data using the Smote sampling approach, implementing the Random Forest method, and testing performance based on accuracy, sensitivity, and specificity. Lung cancer data is obtained from the Kaggle site, which has 276 data instances. Before class balancing using SMOTE, there were 38 negative and 238 positive cancer. Lung cancer data that has not been balanced, then performed data balancing using Smote so that the distribution of each class becomes 238 instancesthe distribution of each class before and after Smote can be seen in Table 2. The next process is to classify lung cancer data using Random Forests. The classification results of the Random Forest method without Smote can be seen in Figure 2, while the results of the combination of Smote and Random Forest are shown in Figure 3.
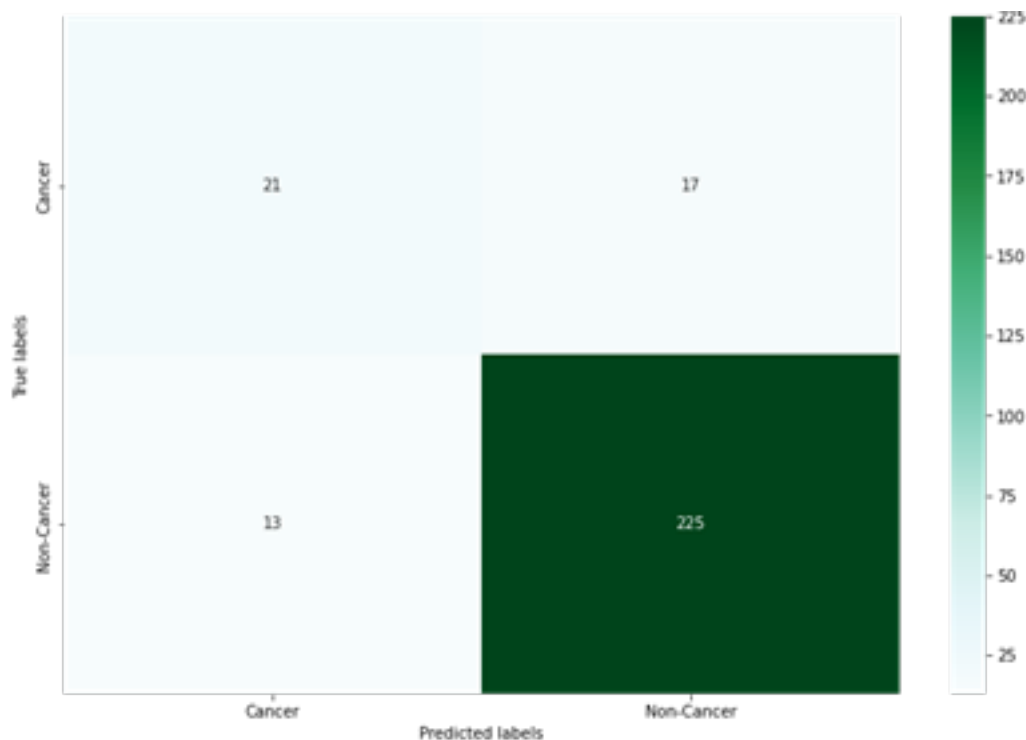

Figure 2. Confusion Matrix Results on Random Method Without Smote

Figure 3. Confusion Matrix Result of Random Forest Method with Smote

Table 2. Class Ratio of Lung Cancer Dataset

Sampling Methods	Number of Instances	
	Negative	Positive
Original	38	238
SMOTE	238	238

In Figure 3, the Random Forest method without Smote correctly predicts the cancer class in as many as 21 instances out of 38 data. In comparison, the non-cancer class is correctly predicted in as many as 225 instances out of 238 data. The Random Forest method with Smote correctly predicts the cancer class in as many as 225 instances out of 238 data, while the non-cancer class correctly predicted as many as 223 instances out of 238 data. The Random Forest method without Smote obtained an accuracy of 89.1%, a sensitivity of 55%, and a specificity of 94.5%. At the same time, the Smote and Random Forest methods combination obtained an accuracy of 94.1%, a sensitivity of 94.5, and a specificity of 93.7%. The use of Smote can improve the performance of the Random Forest classification method based on accuracy and sensitivity. In accuracy, there was an increase of 5%, and the sensitivity of the increase was 39%; this was reinforced by research [15] [16].

4.   CONCLUSION

Based on the test results, the combination of the Smote method with Random Forest gets an accuracy of 94.1%, sensitivity of 94.5%, and specificity of 93.7%. In contrast, without Smote, it gets an accuracy of 89.1%, sensitivity of 55%, and specificity of 94.5%. The Smote and Random Forest combination increased accuracy by 5% and sensitivity by 39% compared to without Smote. This result indicates that using Smote can increase the accuracy and sensitivity of the Random Forest method in lung cancer prediction. Future research suggestions are hybrid samplings such as Smote-Tomek Link or Smote-ENN to solve the problem of unbalanced data in lung cancer to improve the performance of the classification method used better than Smote individually.

5. ACKNOWLEDGEMENTS

We would like to thank those who have supported this research.

6. DECLARATIONS

REFERENCES

[1] F. A. Hermawati and M. I. Safii, "Sistem Deteksi Keganasan Kanker Paru-Paru pada CT Scan dengan Menggunakan Metode Mask Region-based Convolutional Neural Network (Mask R-CNN)," *Proceeding KONIK (Konferensi Nasional Ilmu Komputer)*, vol. 5, pp. 193–197, 2021.

[2] P. Saha, R. O. Nyarko, P. Lokare, I. Kahwa, P. O. Boateng, and C. Asum, "Effect of Covid-19 in Management of Lung Cancer Disease: A Review," *Asian Journal of Pharmaceutical Research and Development*, vol. 10, no. 3, pp. 58–64, 2022.

[3] A. Bhattacharjee, R. Murugan, and T. Goel, "A hybrid approach for lung cancer diagnosis using optimized random forest classification and K-means visualization algorithm," *Health and Technology*, pp. 1–14, 2022.

[4] E. Wulandari, "Klasifikasi Kanker Paru-Paru Menggunakan Metode Naive Bayes," *International Research on Big-Data and Computer Technology: I-Robot*, vol. 6, no. 2, pp. 20–24, 2022.

[5] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," *Jurnal Infortech*, vol. 2, no. 1, pp. 96–101, 2020.

[6] H. Harafani and H. A. Al-Kautsar, "Meningkatkan Kinerja K-NN Untuk Klasifikasi Kanker Payudara Dengan Forward Selection," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 1, pp. 99–110, 2021.

[7] E. Nemlander, A. Rosenblad, E. Abedi, S. Ekman, J. Hasselström, L. E. Eriksson, and A. C. Carlsson, "Lung cancer prediction using machine learning on data from a symptom e-questionnaire for never smokers, formers smokers and current smokers," *PLoS ONE*, vol. 17, no. 10 October, pp. 1–11, 2022.

[8] G. A. Shanbhag, K. A. Prabhu, N. V. Reddy, and B. A. Rao, "Prediction of Lung Cancer using Ensemble Classifiers," in *Journal of Physics: Conference Series*, vol. 2161, no. 1, 2022, pp. 1–11.

[9] A. Helisa, T. H. Saragih, I. Budiman, F. Indriani, and D. Kartini, "Prediction of Post-Operative Survival Expectancy in Thoracic Lung Cancer Surgery Using Extreme Learning Machine and SMOTE," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 2, pp. 239–249, 2023.

[10] Hairani, M. N. Abdillah, and M. Innuddin, "An Expert System for Diagnosis of Rheumatic Disease Types Using Forward Chaining Inference and Certainty Factor Method," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, 2019, pp. 104–109.

[11] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *International Journal on Informatics Visualization*, vol. 7, no. 1, pp. 258–264, 2023.

[12] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, apr 2020. [Online]. Available: https://jtsiskom.undip.ac.id/index.php/jtsiskom/article/view/13544

[13] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022.

[14] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[15] H. Hairani, "Peningkatan Kinerja Metode SVM Menggunakan Metode KNN Imputasi dan K-Means-Smote untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 4, pp. 713–718, jul 2021. [Online]. Available: https://jtiik.ub.ac.id/index.php/jtiik/article/view/3428

[16] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, jun 2020. [Online]. Available: https://doi.org/10.30534/ijatcse/2020/104932020http://www.warse.org/IJATCSE/static/pdf/file/ijatcse104932020.pdf

[This page intentionally left blank.]