JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 4, No. 2. March 2022 DOI: https://doi.org/10.34288/jri.v4i2.385
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

THE EFFECTIVENESS ANALYSIS OF RANDOM FOREST ALGORITHMS WITH SMOTE TECHNIQUE IN PREDICTING LUNG CANCER RISK

Ita Yulianti[1], Ami Rahmawati[2*)], Tati Mardiana[3]

Sistem Informasi Akuntansi Kampus Kota Sukabumi
Universitas Bina Sarana Informatika
ita.iyi@bsi.ac.id[1]

Sistem Informasi[2*)], Sains Data[3]
Universitas Nusa Mandiri
ami.amv@nusamandiri.ac.id[2*)], tati.ttm@nusamandiri.ac.id[3]

(*) Corresponding Author

Abstrak

Jika dibandingkan dengan jenis kanker lainnya, sebagian besar penduduk penderita kanker meninggal karena kanker paru-paru. Seseorang perlu melakukan tes skrining melalui rontgen, CT scan, dan MRI untuk mendeteksi penyakitnya. Namun, sebelum melakukan proses tersebut, dokter biasanya akan melakukan anamnesis dan pemeriksaan fisik terlebih dahulu untuk mempelajari gejala dan kemungkinan faktor risiko kanker paru-paru. Kumpulan data kanker paru memiliki ketidakseimbangan kelas sehingga mempengaruhi kinerja algoritma random forest dalam memprediksi risiko kanker paru. Penelitian ini bertujuan untuk menerapkan teknik SMOTE untuk meningkatkan kinerja algoritma random forest dalam memprediksi risiko kanker paru. Pada penelitian ini pengolahan dan analisis data menggunakan bahasa pemrograman Python. Hasil pengujian menunjukkan nilai akurasi sebesar 88% dengan nilai AUC sebesar 0,93. Saat menggunakan algoritma random forest untuk memperkirakan risiko kanker paru-paru, teknik SMOTE berguna dalam menangani ketidakseimbangan kelas dalam kumpulan data.

Kata kunci: Kanker Paru-Paru, Python, Random Forest, SMOTE

Abstract

When compared with other types of cancer, most of the population with cancer die from lung cancer.A person needs to do a screening test through X-rays, CT scans, and MRI to detect the disease. However, before carrying out the process, the doctor will ordinarily investigate a medical history and physical examination first to study the symptoms and possible risk factors for lung cancer. The lung cancer data set has a class imbalance that affects the performance of the random forest algorithm in predicting the risk of lung cancer. This study aims to employ the SMOTE technique to the random forest algorithm to increase accuracy in predicting lung cancer risk. In this research, data processing and analysis use the Python programming language. The test results show an accuracy value of 88% with an AUC value of 0.93. When employing the random forest method to forecast lung cancer risk, the SMOTE technique is useful in dealing with class imbalances in the data set.

Keywords: Lung Cancer, Python, Random Forest, SMOTE

INTRODUCTION

Cancer is one of the leading causes of significant morbidity and mortality worldwide (Sofia & Tahlil, 2018). In 2018, the number of deaths from cancer reached 9.6 million people. Lung, prostate, colorectal, stomach, liver, and breast cancers are the biggest causes of cancer deaths every year (WHO, 2022). When compared with other types of cancer, most of the population with cancer died from lung cancer (Rattan et al., 2018). According to WHO data from 2020, lung cancer is the leading cause of death, with 1.80 million cases (WHO, 2021).

In Indonesia itself, the majority of lung cancer sufferers are male due to smoking habits which can make the risk of getting cancer higher (Bulan et al., 2017). Lung cancer is difficult to detect because many cases of this cancer appear and show symptoms when its development has reached a certain stage (Makaju et al., 2018). Screening tests through X-rays, CT scans, and MRIs can detect the

disease in the lungs (Kurnia et al., 2016). However, before carrying out this process, the doctor will usually do a medical history and physical examination first to study the symptoms and possible risk factors for lung cancer (American Cancer Society, 2022).

In line with the development of science and technology today, various fields ranging from education, economics, social, health, and many others are increasingly taking advantage of the role of technology, especially in the decision-making process. The many applications of this technology certainly encourage researchers to facilitate human work, for example in the health sector, namely to predict certain diseases using data mining techniques. This technique has many methods/algorithms that can be used, one of which is a random forest which is a modified algorithm from the decision tree. This approach has the advantage of being able to process a random selection of features in image datasets or in the form of attributes/parameters, resulting in a low error rate (Sari et al., 2020).

The random forest algorithm is generally widely used in previous studies including to measure the severity of disease in apple leaf images (Ratnawati & Sulistyaningrum, 2019). In addition, this algorithm is also better than the xgboost algorithm in overcoming the imbalance class in the classification of hepatitis C disease levels (Syukron et al., 2020). Therefore, the purpose of this study was to analyze and measure the effectiveness of the random forest algorithm in predicting the risk of lung cancer through a physical examination in the form of perceived symptoms and habits or lifestyles. Patients with lung cancer usually have symptoms including shortness of breath, chest pain that does not improve, cough with blood, changes in the shape of the fingers, difficulty swallowing, and weight loss (Syifa et al., 2016). Not only that but in this study, the RF algorithm will also be integrated with the SMOTE technique to overcome if there is a class imbalance in the dataset used so that it can improve the performance of the algorithm and the predicted outcome obtained is optimal.

This research is expected to provide novelty in the form of modeling that can be developed into a system that can predict the disease. Although not all lung cancer can be prevented, knowing the early detection of the risk of this disease, allows a person to immediately carry out a follow-up examination (screening) and be treated by a doctor so that it can reduce the risk of lung cancer.

RESEARCH METHODS

The following is an overview of the research framework used to describe all the stages carried out in this research:

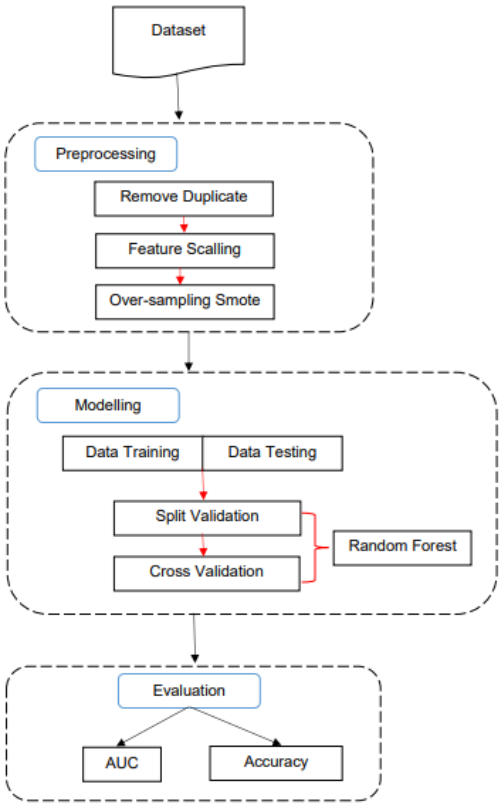

Figure 1. Research Methode framework

The figure above describes the stages or steps taken to achieve the results of this research. This stage begins with selecting a dataset according to the topic raised, namely lung cancer. The next stage is preprocessing which includes removing duplicates, feature scaling and over-sampling smote. In the preprocessing stage, the dataset is processed first before entering the modeling stage. This is done so that the dataset used does not have problems that can complicate the processing. Then move on to the modeling stage, the dataset is divided into training data and testing data with a ratio of 80:20 using split validation combined with cross validation using the tested algorithm, namely random forest. Finally, the results of the algorithm test are validated with the help of a confusion matrix which gives output in the form of accuracy values and AUC graphs.

Data Collection

The dataset in this research is a secondary dataset obtained from the Kaggle Repository with CSV format which has a total of 16 attributes including the target class.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i2.385
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

Preprocessing

In this research, there are three stages of preprocessing the data used, namely, removing duplicates, feature scaling and over-sampling smote. The purpose of removing duplicates is to prevent duplication of data by filtering and deleting the same data in the dataset (Hendra & Fitriyani, 2021). Meanwhile, feature scaling is carried out to examine the diversity of values that occur in each variable and to balance the scale so that it has the same range of values (Aripin, 2021).

Class balance in the application of classification algorithms is important to pay attention to so that the resulting performance has a good prediction. Therefore, to overcome the class imbalance in the research dataset, a resampling technique is used, namely oversampling, this technique was chosen because it can balance the dataset that is lacking in the minority class without reducing the dataset (Sulistiyono et al., 2021). The oversampling algorithm used is Synthetic Minority Over-sampling Technique (SMOTE), where this algorithm is an algorithm that can be called similar to random oversampling (ROS) but the difference is that in SMOTE, existing samples are not duplicated randomly but are made with the nearest neighbor concept (Arifiyanti & Wahyuni, 2020). The workings of the SMOTE algorithm are to interpolate the original data and then enter the artificial data generated in the minority class so that the data obtained varies (Indrawati, 2021). In general, the function of this algorithm is written with SMOTE (X, N, k), where X is the minority data, while N is the percentage of the total instances to be created, and k is the number of closest instances of the instance being searched for using the Euclidean distance formula as following (Sulistiyono et al., 2021):

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_n - Y_n)^2} \quad \text{.....(1)}$$

Modeling

Random Forest was first introduced by Breiman in 2021, which is one of the decision tree-based machine learning methods that is often used because it has the advantage of having high dimensions with a faster process that functions, especially on subset features (Ardiningtyas & Paulina Heruningsih Prima, 2021). This algorithm can be explained as a non-parametric model with the concept of combining learning models to improve performance in both regression and classification cases (Aripin, 2021).

There are three steps in the development of the random forest algorithm, namely (Religia et al., 2021):

1. Set sampling is part of training k,
2. Making each decision tree model, and
3. Collection of k trees into the RF Model

The random forest algorithm is a development method from the CART decision tree, so it is not surprising that in building the decision tree the CART method is used which begins by finding the entropy value first (Aripin, 2021) using Equation (2) below:

$$Entropy\ (Y) = -\sum ip(Y) log_2 p(Y) \quad \text{..... (2)}$$

Then proceed with calculating the information gain using Eq. (3)

$$Information\ Gain\ (Y, a) = Entropy(Y) - \sum values(a) \frac{|Yv|}{|Ya|} (Entropy(Yv)) \quad \text{..... (3)}$$

RESULTS AND DISCUSSION

Dataset

Collecting data that is processed in this study is public data obtained from the Kaggle website. The dataset was collected from the Lung Cancer Prediction System's online site. The following is a description of the lung cancer dataset as shown in Table 1.

Table 1. Lung Cancer Dataset Description

Lung Cancer Dataset	Number of Attributes	Number of Instances
	16	309

From the data obtained, there are attributes, namely: gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, chest pain, and lung cancer. An explanation of each attribute will be described in Table 2.

Table 2. Description of Data Attributes

Attributes	Value
Gender	M (Male), F (Female)
Age	21 – 87
Smoking	Yes = 2, No = 1
Yellow Fingers	Yes = 2, No = 1
Anxiety	Yes = 2, No = 1
Peer Pressure	Yes = 2, No = 1
Chronic Disease	Yes = 2, No = 1
Fatigue	Yes = 2, No = 1
Allergy	Yes = 2, No = 1
Wheezing	Yes = 2, No = 1
Alcohol	Yes = 2, No = 1

Coughing	Yes = 2, No = 1
Shortness Of Breath	Yes = 2, No = 1
Swallowing Difficulty	Yes = 2, No = 1
Chest Pain	Yes = 2, No = 1
Class Lung Cancer	Yes, No

Data Exploration and Visualization

Data exploration and visualization are carried out to obtain information and understand the attributes in the dataset. Exploration and visualization of several attributes are presented in graphical form as follows:

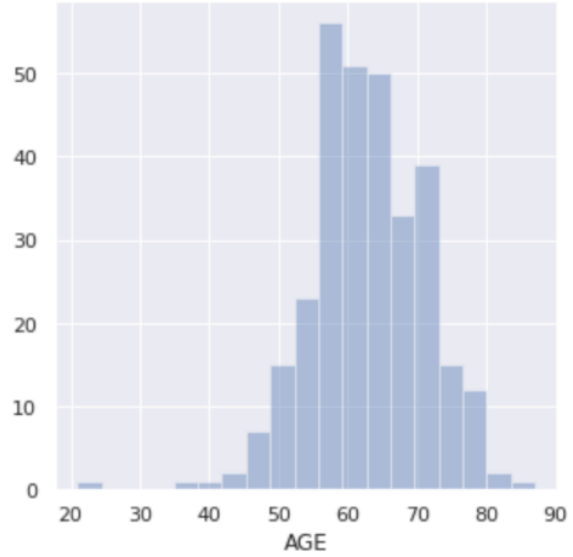

Figure 2. Age Attribute

The results in the first graph (See Figure 2.) show that the age attribute that dominates people at risk for lung cancer is the age of 60-70 years.

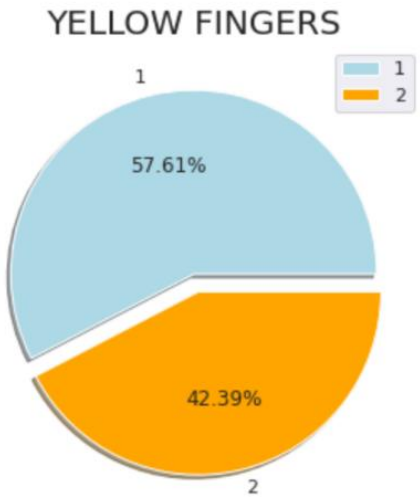

Figure 3. Yellow Fingers Attribute

Then, the second result is in the form of a pie chart that shows the percentage of yellowing fingers symptoms experienced by people who are at risk of lung cancer by 42.39%.

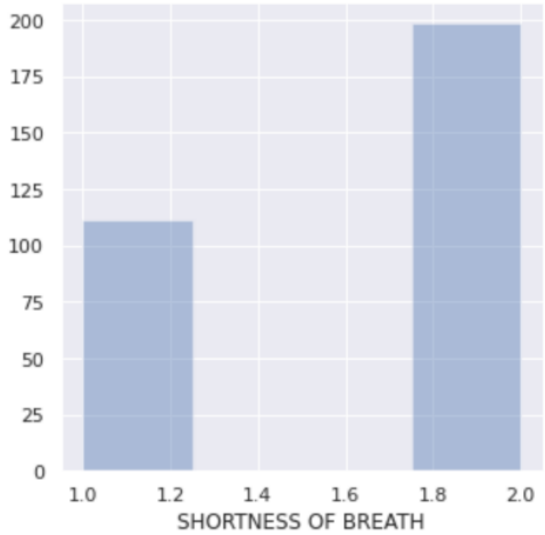

Figure 4. Shortness of Breath Attribute

Based on the results of exploration and visualization of further data, it shows that shortness of breath is a symptom that is often experienced by people who are at risk of lung cancer.

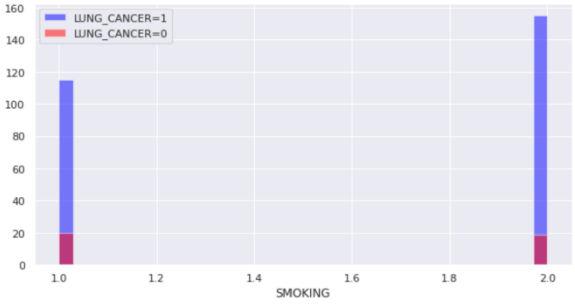

Figure 5. Smoking Attribute

Exploration of data for smoking attributes shows that active smokers are more susceptible to lung cancer.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i2.385
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

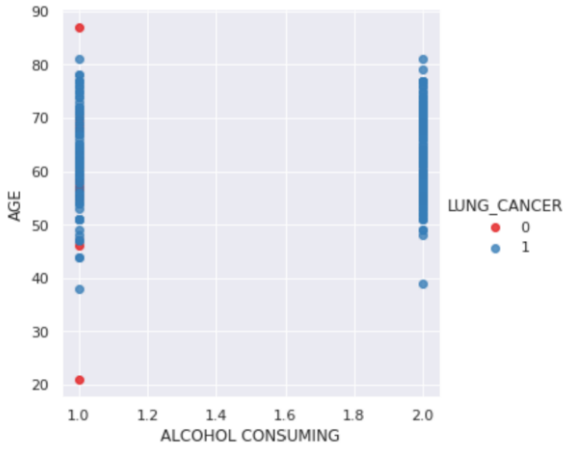

Figure 6. Age dan Alcohol Attribute

The next graph (See Figure 6.) proves that the age range of 50-80 years and consuming alcohol is more at risk of lung cancer.

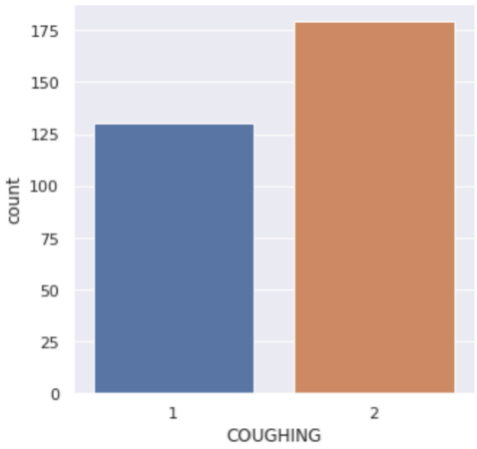

Figure 7. Coughing Attribute

In addition, there is also another attribute that is the most common symptom experienced by people at risk for lung cancer, namely cough (See results of data exploration and visualization in Figure 7.).

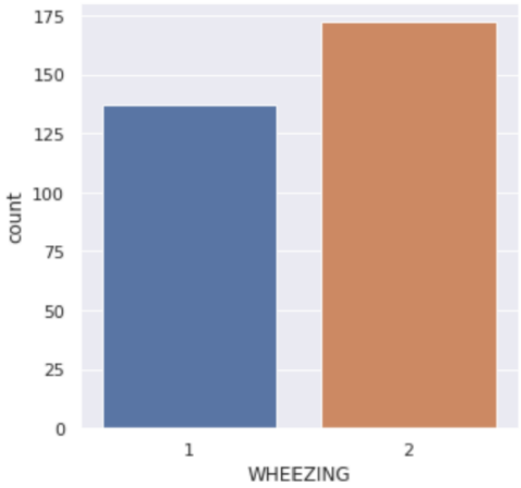

Figure 8. Wheezing Attribute

Finally, the graphic results from the wheezing attribute show that loud, high-frequency breathing sounds are often experienced by people who are at risk for lung cancer.

Modeling Results

a. Remove Duplicate

Removing duplicates is a step taken to remove duplication of data. In the lung cancer dataset, there are 33 duplicates, so removing duplicates is done. The number of tuples in the dataset, which was originally 309, has changed to 276.

b. Feature Scaling

In this research, feature scaling is carried out, namely to make numerical data have the same range of values. The scaling feature used is MinMaxScaler to scale data values into a range.

c. Synthetic Minority Over-sampling Technique (SMOTE)

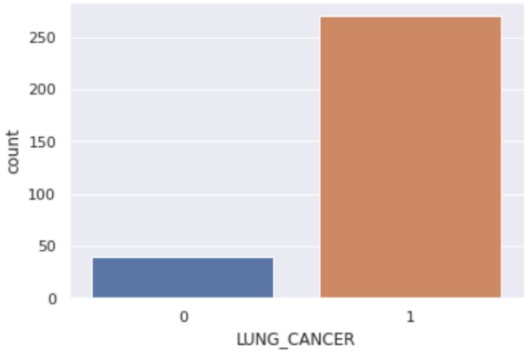

Figure 9. Class Lung Cancer

Based on Figure 9, it can be seen that there is a very large distribution of data between classes, where the Yes class has a sample size of more than

250 data while the No class only has a sample number of fewer than 50 data so that the class is not balanced. To deal with imbalanced classes, in this study, SMOTE was used. SMOTE will select a point from the minority class and calculate the K nearest neighbors for this point.

d. Modeling Results with Random Forest

To determine and test the classification model, a data split was carried out, namely dividing the data into two parts. 80% for training data and 20% for test data. The training data is used to build the model while the test data is used for valid performance evaluation. In addition, cross-validation was also carried out on the test data with a value of K = 10. The results of the tests carried out were to produce values of accuracy (confusion matrix), precision, recall, and AUC. The results of the confusion matrix can be seen in Table 3.

Table 3. Confusion Matrix of RF Algorithm

Classification	Classified as	
	Positive	Negative
Positive	4	5
Negative	2	45

Based on Table 3, there are details on the number of True Positive = 4, False Positive = 5, False Negative = 2, and True Negative = 45. From this data, accuracy, precision, and recall can be calculated. The test result data can be seen in Figure 10 and Table 4.

Table 4. Results of Random Forest Test

Performance	Random Forest
Accuracy	0,88
Precision	0,90
Recall	0,95

```
Classification report Random Forest:
              precision    recall  f1-score   support

           0       0.50      0.22      0.31         9
           1       0.87      0.96      0.91        47

    accuracy                           0.84        56
   macro avg       0.68      0.59      0.61        56
weighted avg       0.81      0.84      0.81        56

Accuracy:  0.88
```

Figure 10. Results of Random Forest Test

In addition, the test results for the random forest algorithm can also be seen based on the ROC graph which can be seen through the resulting AUC value of 0.93. This proves that the accuracy of the model is included in the excellent classification category. The resulting ROC graph can be seen in Figure 11.

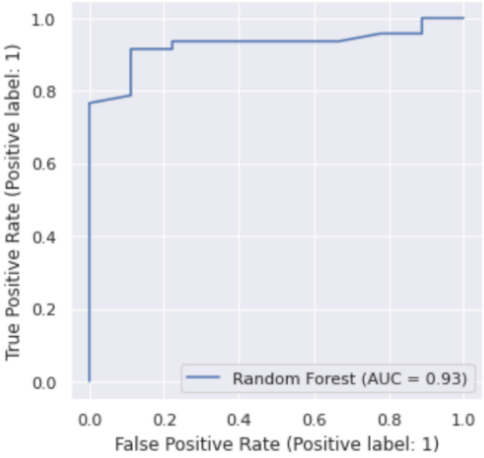

Figure 11. AUC Random Forest

Meanwhile, the precision and recall graphs are visualized through a graph that can be seen in the following figure.

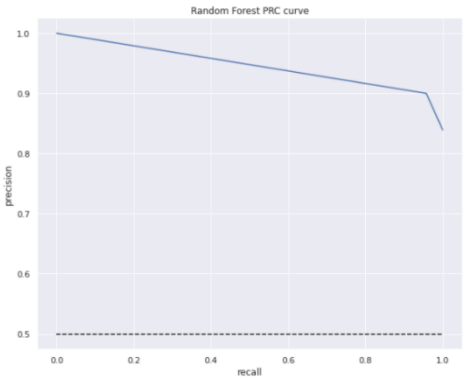

Figure 12. PRC Random Forest

CONCLUSIONS AND SUGGESTIONS

Conclusion

From the research results obtained, the lung cancer dataset has a class imbalance. To overcome the class imbalance, SMOTE (Minority Over Sampling Technique) was used and the random forest algorithm was used for classification testing. Testing the lung cancer dataset yielded an accuracy of 88% with an AUC value of 0.93. While the precision value is 0.90 and the recall value is 0.95. Based on this explanation, it can be concluded that the application of the SMOTE technique to the random forest algorithm can improve the performance of unbalanced data classification, so that the level of effectiveness of the resulting algorithm test becomes more optimal.

Suggestion

For further research, it is hoped that the modeling resulting from this research can be redeveloped by being integrated using optimization algorithms and implemented into a GUI application that can be easily accessed by everyone, as well as adding image datasets so that prediction results are better.

REFERENCES

American Cancer Society. (2022). *Lung Cancer*. American Cancer Society. https://www.cancer.org/cancer/lung-cancer/

Ardiningtyas, Y. E., & Paulina Heruningsih Prima, R. (2021). *ANALISIS BALANCING DATA UNTUK MENINGKATKAN AKURASI DALAM KLASIFIKASI*. 24–28.

Arifiyanti, A. A., & Wahyuni, E. D. (2020). *SMOTE : METODE PENYEIMBANG KELAS PADA KLASIFIKASI DATA MINING*. *XV*, 34–39.

Aripin, H. A. (2021). *Jurnal Informatika dan Komputer (INFOKOM)*. *9*, 30–45.

Bulan, I. A. K. T., Ratnawati, H., Wargasetia, T. L., Bulan, I. A. K. T., Ratnawati, H., & Wargasetia, T. L. (2017). Gambaran pasien kanker paru di rumah sakit Immanuel Bandung periode januari 2013 hingga desember 2014. *Journal of Medicine and Health*, *1*(6), 517–524.

Hendra, A., & Fitriyani. (2021). *Analisis Sentimen Review Halodoc Menggunakan Nai¨ ve Bayes Classifier*. *6*(2), 78–89.

Indrawati, A. (2021). *PENERAPAN TEKNIK KOMBINASI OVERSAMPLING DAN UNDERSAMPLING HYBRID OVERSAMPLING AND UNDERSAMPLING TECHNIQUES TO HANDLING IMBALANCED DATASET*. *4*(1), 38–43. https://doi.org/10.33387/jiko

Kurnia, R., Rahmadewi, R., & Aini, F. (2016). Deteksi Dini Penyakit Paru Dengan Metoda Bayesian Berbasis Android. *National Conference of Applied Engineering, Business and Information Technology, Politeknik Negeri Padang*, 317–323.

Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K., & Elchouemi, A. (2018). Lung Cancer Detection using CT Scan Images. *Procedia Computer Science*, *125*(2009), 107–114. https://doi.org/10.1016/j.procs.2017.12.016

Ratnawati, L., & Sulistyaningrum, D. R. (2019). *Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit*. *8*(2).

Rattan, S., Kaur, S., Kansal, N., & Kaur, J. (2018). An optimized lung cancer classification system for computed tomography images. *2017 4th International Conference on Image Information Processing, ICIIP 2017*, *2018-Janua*, 15–20. https://doi.org/10.1109/ICIIP.2017.8313676

Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). *JURNAL RESTI Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk*. *1*(10), 187–192.

Sari, V. R., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *EDUMATIC : Jurnal Pendidikan Informatika*, *4*(2), 1–9. https://doi.org/10.29408/edumatic.v4i2.2202

Sofia, R., & Tahlil, T. (2018). Pengalaman Pasien Kanker dalam Menghadapi Kemoterapi. *Jurnal Ilmu Keperawatan*, *6*(2), 81–91.

Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *SISTEMASI*, *10*(2), 445. https://doi.org/10.32520/stmsi.v10i2.1303

Syifa, R. A., Adi, K., Fisika, D., & Diponegoro, U. (2016). Analisis Tekstur Citra Mikroskopis Kanker Paru Menggunakan Metode Gray Level Co-Occurance Matrix (Glcm) Dan Tranformasi Wavelet Dengan Klasifikasi Naive Bayes. *Youngster Physics Journal*, *5*(4), 457–462.

Syukron, M. ;, Santoso, R., & Widiharih, T. (2020). ISSN: ISSN: 2339-2541JURNAL GAUSSIAN, Volume9, Nomor 3, Tahun 2020, Halaman 227-236Online di: https://ejournal3.undip.ac.id/index.php/gaussian/PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IM. *Jurnal Gaussian*, *9*(3), 227–236.

WHO. (2021). *Cancer*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer

WHO. (2022). *Cancer*. World Health Organization. https://www.who.int/health-topics/cancer#tab=tab_1