Using Data Sampling Technique for Improving Classification of Covid-19 and Lung Diseases

A R Purnajaya* and F D Hanggara
Universal University, Indonesia

{rezkipurnajaya@gmail.com}

Abstract. The Covid-19 is a virus that has spread around the world and can cause infected respiratory tracts to die. One solution to this problem is to classify the Covid-19 chest X-ray. Among the challenges in this area is improving the classification performance of Covid-19 chest X-rays. Covid-19 chest X-ray and other lung disease chest X-rays have similar colors and patterns, which makes the classification performance not optimal. As a solution to this problem, this study used chest Covid-19 X-ray data and 12 types of other lung diseases chest X-ray data to improve classification performance by applying data sampling techniques. Data sampling techniques included Random Undersampling (RUS), Random Oversampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), and Tomek Link (T-Link) will be evaluated. This study uses Support Vector Machines (SVM) to classify data and evaluation is based on the highest Area Under Curve (AUC) value and accuracy value. ROS found to be the best data sampling technique with an average increase in AUC and accuracy for all datasets of 31.5% and 3.4%, respectively. As a result, the ROS technique helps classify COVID-19 and other lung diseases more accurately.

Keywords: Lung Diseases; SMOTE; Random Oversampling

1 Introduction

On December 31, 2019, the World Health Organization (WHO) China Country Office reported a case of pneumonia of unknown etiology in Wuhan City, Hubei Province, China. On January 7, 2020, China identified pneumonia of unknown etiology as a new type of coronavirus or Covid-19 [1]. The rapid spread of Covid-19 acute respiratory syndrome is causing a worldwide pandemic. Therefore, many countries have implemented preventative measures to control community spread, such as social distancing, restrictions on non-urgent medical care, and the closing of non-essential businesses. Covid-19 continues to spread despite these efforts, posing a global public health crisis [2].

Due to these challenges, a solution is required to diagnose Covid-19-infected patients by predicting whether they are safe or infected. Chest X-ray data can be used as input data for data classification in one implementation of the solution. A similar implementation was carried out by Bergtholdt *et al.* (2016) in the manufacture of a lung nodule detection system with classification. The Area Under Curve (AUC) value obtained is 90% [3]. Classification of data in this study will be carried out using the Support Vector Machines (SVM) model. Based on research comparing data classification models, the SVM classification method was selected because its results were superior. The results of the research by Harefa and Pratiwi (2016) related to mammogram images show that SVM has an overall accuracy rate of 93.98% compared to the k-Nearest Neighbor accuracy rate of 63.86% [4].

UNISET 2021, December 02, Kuningan, Indonesia

DOI 10.4108/eai.2-12-2021.2320242

Classification of patients infected with Covid-19 against healthy patients will have a high accuracy value compared to patients who are infected with other lung diseases [5]. This is due to bias during the classification of fellow lung diseases because there are similarities in color and pattern in the two chest X-ray images. As a result, there is an error in handling patients who have ordinary lung disease but are diagnosed with Covid-19 or vice versa. Moreover, the Covid-19 case began when an unknown pneumonia case was reported in Wuhan, Hubei Province, China. After analyzing the respiratory samples, experts from the Chinese State Disease Center stated that the pneumonia was caused by Covid-19 [6].

In order to overcome these problems, it has been proven that the data sampling technique improves data classification. At the time of collecting chest X-ray images, it is always found that the number of data classes is different from the number of other data classes. This is known as imbalanced data. If you try to balance the number of data classes manually, you will lose data which may be important. However, if the imbalanced data is tried to continue to be classified, it will cause the average misclassification to be higher in the minor class data. The data sampling technique changes the imbalanced data by different procedures to produce a balanced distribution of data. Balanced data can improve the overall performance of the classification process compared to unprocessed data [7].

For this reason, this research was formulated to gain knowledge of what data sampling techniques can improve the classification performance of Covid-19 with other lung diseases. The evaluation results from the comparison of data sampling techniques will be measured based on the AUC and the level of accuracy as an additional parameter.

2 Methodology

2.1 Preparation of Chest X-ray Data

This stage is carried out to prepare the dataset used for data sampling and classification. This study uses chest X-ray data infected with Covid-19 and 12 chest X-ray data infected with other lung diseases. Covid-19 and Viral Pneumonia chest X-ray was sourced from Chowdhury *et al.* (2020) and Rahman *et al.* (2020) [8]-[9], Tuberculosis chest X-ray was sourced from Rahman *et al.* (2020) [10], then Atelectasis, Consolidation, Effusion, Emphysema, Infiltration, Nodule, Pleural Thickening, Pneumothorax, Edema, and Mass chest X-ray was sourced from Wang *et al.* (2017) [11]. Description of chest X-ray can be seen in Table 1.

Table 1. Chest X-Ray Data of Covid-19 and Lung Disease

Covid-19 (3616 images)	RNA viruses with a particle size of 120-160 nm. The coronavirus that causes COVID-19 is included in the beta coronavirus genus [12].
Tuberculosis (700 images)	Direct infectious disease caused by the bacteria *Mycobacterium Tuberculosis*. *Mycobacterium Tuberculosis* bacteria can live mainly in the lungs and can be transmitted to various other body organs through the blood such as the digestive tract [10].
Viral Pneumonia (1345 images)	Pneumonia is an inflammatory process in which there is consolidation caused by the filling of the alveoli cavity by exudate. When consolidation occurs, it prevents gas exchange, as well as blood flow around the alveoli, which does not function optimally [13].
Atelectasis (508 images)	Atelectasis occurs when the lungs are unable to fully expand [13]. This condition results in a decrease in the surface area available for the diffusion process and a reduced respiratory rate [14].
Consolidation (271	Pulmonary consolidation is a term that describes when a lung X-ray

images)	appears white or cloudy [15].
Effusion (644 images)	Effusion is a condition characterized by the accumulation of fluid in the pleural cavity on the surface of the visceral and parietal pleura. Additionally to fluid, pleural effusions contain pus and blood [16].
Emphysema (127 images)	Emphysema is an obstructive pulmonary disease that is chronic and progressive, characterized by anatomic abnormalities in the form of widening of the distal air cavities in the terminal bronchioles and damage to the lung parenchyma [17].
Infiltration (967 images)	Pulmonary infiltrates are substances denser than air that persist within the parenchyma of the lung, such as pus, blood, or protein. They are more common in febrile neutropenic patients [18].
Nodule (313 images)	A lung nodule is defined as a “spot” on the lung that is 3 centimeters (approximately 1.5 inches) in diameter or less. If the abnormality seen on X-ray of the lung is larger than 3 centimeters [19].
Pleural Thickening (176 images)	The pleura is a layer that covers the lung tissue. The pleural layer consists of 2 layers, namely the visceral pleural layer which covers and attaches directly to the lung tissue and the parietal pleural layer which adheres to the chest wall [20].
Pneumothorax (271 images)	Pneumothorax is the case where an amount of air in the chest increases markedly and a one-way valve is formed leading to a tension pneumothorax. Unless reversed by effective treatment, this situation can progress and cause death [21].
Edema (118 images)	Edema is the result of an imbalance in the filtration system between the capillary and interstitial spaces. The kidneys play a key role in regulating extracellular fluid volume by adjusting sodium and water excretion [22].
Mass (284 images)	A pulmonary mass is any area of pulmonary opacification that measures more than 30 mm. Spots smaller than 3 cm in diameter are considered lung nodules. The commonest cause of a pulmonary mass is lung cancer [23].

Then, all chest X-ray data were converted to 50 x 50 pixels. The test in this study used 12 datasets in which there were two data classes, namely the chest X-ray data class of patients infected with Covid-19 and the chest X-ray data class of patients infected with lung diseases other than Covid-19. The number of chest X-ray data in the 12 datasets is shown in Table 2. Table 2 shows that the number of data classes in this dataset is imbalanced, so data sampling will be used to balance this dataset.

After the feature extraction is done, the next step is to define a class to each category of image data. The definition of classes to the image data is adjusted to the imbalanced dataset where 0 is for the major class and 1 is for the minor class. The major class is in the Covid-19 chest X-ray data because it has large numbers than other lung disease data. In other words, the Covid-19 X-ray data class is 0 and the other lung disease X-ray data class is 1.

Table 2. Total Class Data on Imbalanced Dataset

Imbalanced Dataset	Covid-19 data	Lung Disease data	Imbalanced Dataset	Covid-19 data	Lung Disease data
Covid-19 & Tuberculosis	3616	700	Covid-19 & Infiltration	3616	967
Covid-19 & Viral Pneumonia	3616	1345	Covid-19 & Nodule	3616	313
Covid-19 & Atelectasis	3616	508	Covid-19 & Pleural	3616	176

			Thickening		
Covid-19 & Consolidation	3616	271	Covid-19 & Pneumothorax	3616	271
Covid-19 & Effusion	3616	644	Covid-19 & Edema	3616	118
Covid-19 & Emphysema	3616	127	Covid-19 & Mass	3616	284

2.2. Sampling Data

The explanation of the imbalanced data and the 5 sampling data used are as follows:

- Data Imbalanced
  The initial chest X-ray image data in Table 2 shows the existence of unbalanced data where the amount of sample data is not balanced. An imbalanced dataset occurs when there are fewer members of one class than the other, resulting in an unbalanced class distribution. As a result, minor classes are ignored and classification performance is reduced [7].
- Random Undersampling (RUS)
  RUS is a non-heuristic method that balances the distribution of classes through random deletion of major classes to obtain a balanced set of instances [7].
- Random Oversampling (ROS)
  ROS is a non-heuristic method that balances the distribution of classes by randomly adding data to minor classes [7].
- Combination of Over-Undersampling (COUS)
  COUS is a combined technique algorithm by over-sampling the minor class and under-sampling the major class until both classes have the same number of samples [24].
- Synthetic Minority Over-sampling Technique (SMOTE)
  The SMOTE technique performs oversampling based on k-NN (k-Nearest Neighbors) of the minor class. SMOTE aims to enrich the minor class boundary by creating artificial examples in the minor class rather than adding to the existing examples to avoid overfitting problems [25].
- Tomek Link (T-Link)
  In the dataset, some instances are the nearest neighbor data and are in a different class. This T-Link technique searches for these instances using 1-NN (One-Nearest-Neighbor) in the dataset. To overcome imbalanced, the instance major class is deleted [26].

2.3. Training Data and Test Data

This stage is carried out to create training data and test data used in data classification. Training data or training data is the data used to create a classification model. Test data or testing data is data that will be used to evaluate the classification model and data sampling technique. This study divides the proportion of training data and test data as much as 4:1 from each imbalanced dataset and dataset results from every five samplings. Furthermore, the actual data and the classification results of each dataset are stored into a separate Table for evaluation of the sampling technique.

2.4. Data Classification

Classification of data is carried out after sampling the data. The method chosen for data classification in this study is Support Vector Machines (SVM). SVM is a classification method that is widely used because classification accuracy is strongly influenced by setting kernel parameters and feature selection. SVM groups data by looking for a suitable hyperplane that can separate data based on the highest margin [27].

2.5. Evaluation of Sampling Techniques

Evaluation is done by calculating the value of accuracy and AUC for each data sampling technique. To measure these criteria, predictions will be made using imbalanced data and those that have gone through a sampling process. The results obtained from comparing the prediction results with the actual is a confusion matrix. The results of the confusion matrix were used to calculate the accuracy, specificity, sensitivity, and AUC [28]. The equations that will be used for these criteria are as follows:

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)}$$

$$AUC = \frac{Sensitivity + Specificity}{2}$$

$$Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{(Length\ Positive + Length\ Negative)}$$

3 Result and Discussion

By using the SVM algorithm on imbalanced data and 5 sampling data shown in Table 3, a comparison of performance classification accuracy (Acc) and AUC has been obtained. The Viral Pneumonia and Tuberculosis dataset has been able to classify the Covid-19 dataset well, with the AUC value and accuracy exceeding 90%, while the Infiltration, Effusion, Mass, Nodule, Pneumothorax, Atelectasis, Pleural Thickening, Consolidation, Edema, and Emphysema datasets are still in the category classification performance that is not optimal.

Table 3. Classification Performance on Imbalanced Data and 5 Sampling Data

Dataset	**Sampling Data**	**AUC (%)**	**Acc (%)**	**Dataset**	**Sampling Data**	**AUC (%)**	**Acc (%)**	**Dataset**	**Sampling Data**	**AUC (%)**	**Acc (%)**
Covid-19 & Tuberculosis	Imbalanced	91,6	96,6	**Covid-19 & Effusion**	Imbalanced	74,9	91,2	**Covid-19 & Pleural Thickening**	Imbalanced	54,2	97,1
	RUS	95,0	95,0		RUS	80,9	80,6		RUS	73,4	73,2
	ROS	99,2	99,2		ROS	96,0	96,0		ROS	99,8	99,8
	COUS	98,6	98,6		COUS	96,0	96,0		COUS	76,2	95,3

	SMOTE	99,0	99,1		SMOTE	92,8	93,0		SMOTE	92,2	92,7
	T-Link	93,3	97,6		T-Link	75,1	90,1		T-Link	56,1	95,1
Covid-19 & Viral Pneumonia	Imbalanced	97,7	98,3	**Covid-19 & Emphysema**	Imbalanced	50,0	96,7	**Covid-19 & Pneumothorax**	Imbalanced	61,7	95,4
	RUS	98,0	98,0		RUS	81,4	82,4		RUS	81,3	81,7
	ROS	99,4	99,4		ROS	99,6	99,6		ROS	99,2	99,2
	COUS	99,1	99,1		COUS	95,7	97,3		COUS	83,1	96,1
	SMOTE	99,9	99,9		SMOTE	93,4	93,3		SMOTE	91,3	91,8
	T-Link	97,5	98,3		T-Link	50,0	95,5		T-Link	63,9	95,7
Covid-19 & Atelectasis	Imbalanced	61,6	89,5	**Covid-19 & Infiltration**	Imbalanced	78,7	89,9	**Covid-19 & Edema**	Imbalanced	50,0	97,2
	RUS	77,0	77,0		RUS	81,4	81,4		RUS	78,3	79,2
	ROS	96,7	96,8		ROS	94,3	94,3		ROS	99,3	99,3
	COUS	94,6	94,5		COUS	71,1	94,8		COUS	92,2	97,1
	SMOTE	92,3	92,3		SMOTE	93,9	94,0		SMOTE	94,7	95,2
	T-Link	67,9	92,1		T-Link	81,1	90,6		T-Link	50,0	96,2
Covid-19 & Consolidation	Imbalanced	51,4	95,4	**Covid-19 & Nodule**	Imbalanced	63,8	94,7	**Covid-19 & Mass**	Imbalanced	64,8	95,1
	RUS	79,1	79,1		RUS	77,9	77,0		RUS	76,7	76,3
	ROS	98,3	98,3		ROS	98,0	98,0		ROS	98,5	98,5
	COUS	97,2	97,3		COUS	74,7	95,0		COUS	79,2	93,7
	SMOTE	93,7	93,7		SMOTE	93,1	93,4		SMOTE	94,4	94,5
	T-Link	56,3	95,3		T-Link	68,2	95,5		T-Link	66,9	94,5

Table 4. Increase in Accuracy Value and AUC for each Data Sampling

Dataset	**RUS**		**ROS**		**COUS**		**SMOTE**		**T-Link**	
	AUC (%)	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**	**AUC (%)**	**Acc (%)**
Covid-19 & Tuberculosis	3,4	-1,6	7,6	2,5	45,8	1,8	7,5	2,4	1,8	1,0
Covid-19 & Viral Pneumonia	0,3	-0,3	1,7	1,1	45,7	0,7	2,2	1,6	-0,1	0,0
Covid-19 & Atelectasis	15,4	-12,5	35,2	7,3	42,2	-0,1	30,7	2,8	6,4	2,6
Covid-19 &	27,7	-16,3	46,9	2,9	33,0	5,1	42,3	-1,8	4,9	-0,2

Consolidation										
Covid-19 & Effusion	6,0	-10,6	21,1	4,8	22,0	-1,8	17,9	1,8	0,2	-1,1
Covid-19 & Emphysema	31,4	-14,3	49,6	2,9	21,4	0,8	43,4	-3,4	0,0	-1,2
Covid-19 & Infiltration	2,7	-8,5	15,6	4,5	21,1	4,8	15,2	4,2	2,4	0,7
Covid-19 & Nodule	14,1	-17,7	34,2	3,3	14,4	-1,4	29,3	-1,3	4,4	0,8
Covid-19 & Pleural Thickening	19,2	-23,9	45,6	2,7	10,9	0,4	38,0	-4,4	1,9	-2,0
Covid-19 & Pneumothorax	19,6	-13,7	37,5	3,8	7,0	2,0	29,6	-3,5	2,2	0,4
Covid-19 & Edema	28,3	-18,0	49,3	2,1	1,4	0,8	44,7	-2,0	0,0	-1,0
Covid-19 & Mass	11,9	-18,8	33,7	3,4	-7,6	4,9	29,5	-0,7	2,1	-0,6
Average	15,0	-13,0	31,5	3,4	21,4	1,5	27,5	-0,3	2,2	0,0

According to Table 4, the RUS can significantly increase the AUC value for Covid-19 & Emphysema by 31.4% when the dataset includes classification performance that is not optimal. Although the RUS could not increase the accuracy value, it decreased accuracy to -0.3% in the Covid-19 & Viral Pneumonia dataset and -23.9% in the Covid-19 & Pleural Thickening dataset. In 12 datasets using ROS, the average increase in AUC and accuracy was 15% and -13%, respectively.

Using ROS increases AUC value in Covid-19 & Emphysema and Covid-19 & Edema datasets more significantly, as both datasets have the lowest AUC values in the imbalanced data classification. Additionally, the ROS can increase the accuracy of imbalanced data classification by at least 1.1% in the Covid-19 & Viral Pneumonia dataset and 7.3% in the Covid-19 & Atelectasis dataset. The average increase in AUC and accuracy in 12 datasets using ROS was 31.5% and 3.4%, respectively.

The COUS method can increase AUC value more significantly in the Covid-19 & Consolidation, Covid-19 & Emphysema, and Covid-19 & Edema datasets. COUS accuracy has decreased in the Covid-19 & Edema, Covid-19 & Pleural Thickening, and Covid-19 & Mass datasets. The accuracy value for imbalanced data classification can be increased by COUS by a minimum of 0.7% in the Covid-19 & Emphysema dataset and a maximum of 4.9% in the Covid-19 & Infiltration dataset. In 12 datasets using COUS, AUC increased by 21.4% and accuracy increased by 1.5% on average.

SMOTE can significantly increase the AUC value in the Covid-19 & Edema, Covid-19 & Emphysema, Covid-19 & Consolidation, Covid-19 & Pleural Thickening, and Covid-19 & Atelectasis datasets that are classified as data imbalanced with the lowest AUC. SMOTE has decreased accuracy in many datasets. The Covid-19 & Infiltration dataset had a significant increase in accuracy by 4.2%. The average increase in AUC and accuracy for the 12 datasets using the SMOTE was 27.5% and -0.3%, respectively. Tomek Link sampling did not significantly increase accuracy and AUC in 13 datasets using the Tomek Link sampling technique. The average increase in AUC and accuracy is 2.2% and 0%, respectively.

4 Conclusion

Data analysis was performed using imbalanced data and data sampling techniques for chest X-ray data infected with Covid-19 and 12 chest X-ray data infected with other lung diseases. Based on the results of the evaluation, it can be concluded that the ROS technique accurately samples imbalanced data more accurately than other techniques. RUS, ROS, COUS, SMOTE, and T-Link data were sampled, and their average AUC values were 15.0%, 31.5%, 21.4%, 27.5%, and 2.2%, respectively, with ROS having the highest value at 31.5%. ROS can also improve accuracy with imbalanced data by 3.4%. In other words, ROS is the most accurate data sampling technique for improving classification of Covid-19 and Lung Diseases and can reduce misdiagnosis of Covid-19 disease with other lung diseases.

For further research, a larger number of datasets and a larger image size are required in order to identify changes in the evaluation criteria. This should lead to improvements in the performance of data classification. It will also be interesting to compare these 5 sampling techniques with other sampling techniques in order to determine which technique can improving classification of Covid-19 and Lung Diseases.

5 Acknowledgement

Researchers would like to thank the *Kementerian Pendidikan Kebudayaan, Riset, dan Teknologi, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi* for providing research funding assistance that supported our research.

References

[1] K. K. R. Indonesia, “Pedoman Kesiapsiagaan Menghadapi Infeksi Novel Coronavirus (2019-nCov),” *Direkorat Jenderal Pencegahan dan Pengendalian Penyakit*, pp. 0-74, 2020.

[2] J. Segars, Q. Katler, D. B. McQueen et. al, “Prior and Novel Coronaviruses, COVID-19, and Human Reproduction: What Is Known?,” *Fertil and Steril*, vol. 113, no. 6, pp. 1140-1149, 2020.

[3] M. Bergtholdt, R. Wiemker, and T. Klinder, “Pulmonary nodule detection using a cascaded SVM classifier,” *Medical Imaging:Comp-Aid Diag.*, vol. 9785, p. 978513, 2016.

[4] J. Harefa and M. Pratiwi, “Comparison Classifier: Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) In Digital Mammogram Images,” *Juisi.*, vol. 2, no. 2, pp. 35–40, 2016.

[5] A. R. Purnajaya and F. D. Hanggara, “Perbandingan Performa Teknik Sampling Data untuk Klasifikasi Pasien Terinfeksi Covid-19 Menggunakan Rontgen Dada,” *JAIC.*, vol. 5, no. 1, pp. 37-42, 2021.

[6] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, *et al*., “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, p. 012029, Dec. 2017.

[7] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, “Learning from Imbalanced Data Sets," *Cham: Springer International Publishing*, pp. 1- 377, 2018.

[8] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, M. B. I. Reaz, and M. T. Islam, “Can AI help in screening Viral and COVID-19 pneumonia?,” IEEE Access, vol. 8, pp. 132665 - 132676, 2020.

[9] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. A. Maadeed, S. M. Zughaier, M. S. Khan, and Chowdhury, “Exploring the Effect of Image

Enhancement Techniques on COVID-19 Detection using Chest X-ray Images," Computers in Biology and Medicine, vol. 132, p. 1043192, 2020.

[10] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Ayari, and M. E. H. Chowdhury, "Reliable Tuberculosis Detection using Chest X-ray with Deep Learning, Segmentation and Visualization," IEEE Access, vol. 8, pp. 191586 - 191601, 2020.

[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," IEEE CVPR, pp. 2097-2106, 2017.

[12] A. E. Gorbalenya, S. C. Baker, R. S. Baric, R. J. D. Groot, C. Drosten, A. A. Gulyaeva, *et al*, "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2," Nat Microbiol., vol. 5, pp. 536-544, 2020.

[13] I. Somantri, "Asuhan Keperawatan Pada Klien Dengan Gangguan Sistem Pernafasan," Salemba Medika, p. 162, 2009.

[14] E. J. Corwin, "Buku Saku Patofisiologi," EGC, p. 842, 2009.

[15] A. H. Nurarif and H. Kusuma, "APLIKASI Asuhan Keperawatan Berdasarkan Diagnosa Medis & NANDA NIC-NOC," MediAction, p. 350, 2015.

[16] A. S. Wijaya and Y. M. Putri, "KMB 2 Keperawatan Medikal Bedah Keperawatan Dewasa," Nuha Medika, p. 266, 2013.

[17] C. F. Vogelmeier, G. J. Criner, F. J. Martinez, *et al*, "Global Initiative for Chronic Obstructive Pulmonary Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease," Medical Communications Resources, 2017.

[18] W. Emmanuel and C. Kibbler, "Antibiotic and Chemotherapy (Ninth Edition)." 2010.

[19] A. Jusuf, E. Syahruddin, A. Wibawanto, A. Icksan, Juniarti, and S. Endardjo, "Kanker Paru (Kanker Paru Jenis Karsinoma Bukan Sel Kecil)," Perhimpunan Dokter Paru Indonesia, 2016.

[20] R. Meldau, J. Peter, G. Theron, G. Calligaro, B. Allwood, *et al*, "Comparison of same day diagnostic tools including Gene Xpert and unstimulated IFN-γ for the evaluation of pleural tuberculosis: a prospective cohort study," BMC Pulmonary Medicine, vol. 14, pp. 1-10, 2014.

[21] P. Zarogoulidis, I. Kioumis, G. Pitsiou, K. Porpodis, S. Lampaki, *et al*, "Pneumothorax: from definition to diagnosis and treatment," J Thorac Dis., vol. 6, no. 4, pp. S372–S376, 2014.

[22] J. Jones and V. Tatco, "Treatment of edema," American Family Physician, vol. 71, no. 11, pp. 2111-7, 2021.

[23] J. Jones and V. Tatco, "Pulmonary mass," Radiopaedia.org. (accessed on 9 Nov 2021), 2017.

[24] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and data sampling with imbalanced big data," *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 175-183, 2019.

[25] O. Komori and S. Eguchi, "Statistical Methods for Imbalanced Data in Ecological and Biological Studies," *SpringerBriefs in Statistics*, 2019.

[26] A. D. Pozzolo, O. Caelen, S. Waterschoot, and G. Bontempi, "Racing for Unbalanced Methods Selection," *IDEAL 2013*, pp. 24-31, 2013.

[27] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Applied Soft Computing Journal. Elsevier B.V.*, p. 105946, 2020.

[28] A. R. Purnajaya, W. A. Kusuma, and M. K. D. Hardhienata, "Performance Comparison of Data Sampling Techniques to Handle Imbalanced Class on Prediction of Compound-Protein Interaction," *Biogenesis: Jurnal Ilmiah Biologi*, vol. 8, no. 1, pp. 41-48, 2020.