Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 2, No. 1 (2020), pp. 24–32
P-ISSN 2686-5335, E-ISSN 2716-2478
doi: 10.15408/inprime.v2i1.14545

Classification of Tuberculosis and Pneumonia in Human Lung Based on Chest X-Ray Image using Convolutional Neural Network

Muhaza Liebenlito, Yanne Irene, and Abdul Hamid
Department of Mathematics, Faculty of Sciences and Technology,
UIN Syarif Hidayatullah Jakarta, Indonesia
Email: {muhazaliebenlito, yanne.irene@uinjkt.ac.id}, abdul.hamid15@mhs.uinjkt.ac.id

Abstract

In this paper, we use chest x-ray images of Tuberculosis and Pneumonia to diagnose the patient using a convolutional neural network model. We use 4273 images of pneumonia, 1989 images of normal, and 394 images of tuberculosis. The data are divided into 80% as the training set and 20% as the testing set. We do the preprocessing steps to all of our images data, such as resize, convert RGB to grayscale, and Gaussian normalization. On the training dataset, the sampling technique used is undersampling and oversampling to balance each class. The best model was chosen based on the Area under Curve value i.e. the area under the curve of Receiver Operating Characteristics. This method shows that the best model obtains when trains the training dataset using oversampling. The Area under Curve value is 0.99 for tuberculosis and 0.98 for pneumonia. Therefore, this best model succeeds to identify 86% true for tuberculosis and 96% true for pneumonia.

Keywords: chest X-ray images; tuberculosis; pneumonia; convolutional neural network.

Abstrak

Pada penelitian ini memanfaatkan data citra *chest x-ray* penderita penyakit tuberculosis dan pneumonia. Model *convolutional neural network* digunakan untuk membantu mendiagnosis kedua penyakit ini. Data yang digunakan masing-masing sudah dilabeli sebanyak 4273 citra pneumonia, 1989 citra normal dan 394 citra tuberculosis. Data tersebut dibagi menjadi 80% himpunan data latih dan 20% data uji. Himpunan data tersebut telah melalui 3 tahap *prepocessing* yaitu *resize* citra, merubah citra RGB menjadi *grayscale* dan standarisasi gausian pada citra. Pada data latih dilakukan teknik sampling berupa *undersampling* dan *oversampling* data untuk menyeimbangkan data latih antar kelas. Model terbaik dipilih berdasarkan nilai *Area under Curve* yaitu luas daerah di bawah kurva *Receiver Operating Chracteristics*. Hasil menunjukkan bahwa model terbaik dihasilkan ketika dilatih menggunakan data latih hasil oversampling dengan nilai Area under Curve kelas tuberculosis sebesar 0,99 dan nilai Area under Curve kelas pneumonia sebesar 0,98. Oleh karena itu, model terbaik ini mampu mengindentifikasi sebanyak 86% penyakit tuberculosis dan 96% penyakit pneumonia.

Kata Kunci: citra chest X-ray; penyakit infeksi paru; pengolahan citra digital *Convolutional Neural Network*.

1. INTRODUCTION

The lung is one of the important organs in respiration systems. Lung problems commonly suffered from infections caused by viruses, bacteria, or fungi. Based on the World Health Organization (WHO) statistics, tuberculosis cases in children in 2017, an estimated 1 million suffer from tuberculosis and 230,000 of them died [1]. Whereas pneumonia in 2015 accounted for a total of 16% of deaths in children under the age of 5 years [2]. Tuberculosis and pneumonia are included in lung infections. These two diseases are generally characterized by persistent cough symptoms. Moreover, the cough in tuberculosis is accompanied by blood.

The tool or technique to diagnose tuberculosis and pneumonia is the X-raying the patient's chest. This technique is more economic and easy to use [3]. The image data generated from this technique is called a chest x-ray image (CXR). Several studies have been conducted related to identifying lung disease using the machine learning algorithm on CXR data. Schilham et al. [4] used the Gaussian kernel to extract features and characterize lung nodules features (nodules on the lungs as possible signs of lung cancer) and then classify them with the k-nearest neighbors (KNN) algorithm. Their scheme with the best performance succeeds to identify 51% of all nodules on average two false positives per image and increased to 67% for four false positives.

For tuberculosis, Tan et al. [5] suggested the tuberculosis index (TI) based on the features of the texture segmentation of the lung area which were then classified between normal and abnormal with 95 data using a decision tree algorithm and obtained an accuracy of 92.9%. Hwang et al. [6] were the first researcher to suggest the use of Convolutional Neural Network as an automatic system to detect the tuberculosis disease using transfer learning to improve system performance. Using the KIT, MC data set and Shenzhen, they succeed to make the accuracy system until 90.3%.

Research on the combination of several lung diseases was carried out by Bar et al. [7] using the combination of features extracted by the Convolutional Neural Network (CNN) method and the ImageNet architecture. They succeeded in detecting heart enlargement, right pleural effusion, healthy vs. other abnormal diseases with Area under Curve (AUC) values of 93%, 89%, and 79%, respectively. Abivey et al. [8] made their own CNN architecture succeed to defeat the CNN model with GIST, VGG16 and VGG19 architectures in classifying 14 types of lung disease with a total accuracy of 92.4%.

Many problems in the detection of diseases of the lungs use machine learning methods, especially the CNN method. This is conducted because we need the preprocessing process fewer and the accuracy of this model more promising. Therefore, in this study, we used the CNN method to overcome the problem of classification to detect lung diseases (tuberculosis and pneumonia) using CXR data. There are three categories of classification, namely pulmonary tuberculosis, pneumonia, and normal conditions. We use undersampling and oversampling techniques to handle imbalanced data on the training dataset. Confusion matrix and AUC value will use to the metric evaluation of prediction model for both result oversampling and undersampling.

2. METHOD

In this paper, according to [8], we use the hyperparameter of CNN. In Table 1 is the structure of the CNN architecture that will be used where the term "Conv" represents the convolution layer, "Maxpool" represents the pooling layer used, which is maximum pooling, ReLU represents the activation function on each of the hidden layers of CNN and "FC" represents fully connected layer which amounts to three nodes according to the number of classes to be predicted.

Table 1. The architecture of the CNN model.

Layers	Description	Value
Input Layer	Input Images	32×32×1
Hidden 1	Conv1 + ReLU Maxpool1	16 kernel with size 2×2 and stride 2. Size of kernel 2×2.
Hidden 2	Conv2+ ReLU Maxpool2	32 kernel with size 2×2 and stride 2. Size of kernel 2×2.
Hidden 3	Conv3 + ReLU	64 kernel with size 2×2
Classification Layer	FC	3 nodes.

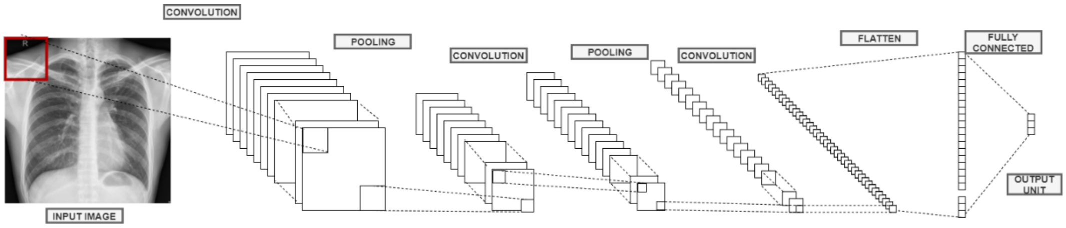

Figure 1. The illustration diagram of the CNN method in this research.

An illustration of the training process from input to output of the CNN model can be seen in Figure 1. In this study, CNN model training will be conducted two times, first for undersampling data results and second with oversampling results data. Each loss model and execution time that produced on the training steps will be taken.

When the model training is done, it will use an optimization algorithm that is Stochastic Gradient Descent (SGD) with a learning rate 0.0001 so that when the training of the value of loss can reach the global minimum well, therefore the maximum epoch that Selected by 2000 epoch. Momentum 0.9 is used to accelerate the model in minimizing its loss value, and a batch size of 50 is chosen for model training to be performed with 50 data at once in a single iteration so that it can trim the training time needed.

We secondary data that is downloaded from the Kaggle where the data for pneumonia in [9] and tuberculosis in [10]. This data is combined data of several diseases, such as:

a. Pneumonia

This data source [11] Containing CXR-as-5863 (JPEG) normal and pneumonia image, this dataset is taken from a group of one to five-year-old children in Guangzhou Women and children's Medical Center, Guangzhou China. This data set was previously through quality control by removing the results of low quality or unreadable scan after that data was diagnosed by two experts [12].

b. Tuberculosis

This tuberculosis data set contains a combination of two sources of data set, such as:

i. *China Set - The Shenzhen set*

This dataset was created by the National Library of Medicine, Maryland, USA in cooperation with Shenzhen No. 3 People's Hospital, Guangdong Medical College, Shenzhen, China [13]. Data is retrieved based on Philips DR Digital Diagnose systems. This data set contains 336 images infected with normal tuberculosis and 326 diseases.

ii. *Montgomery County X-ray Set*

This dataset was created by The National Library of Medicine in collaboration with The Department of Health and Human Services of Montgomery County, MD, USA. Datasets contain 58 CXR images of tuberculosis and 80 normal CXR images [14].

After collecting CXR data, combined each dataset into a folder according to its label, so the amount of data in the class of pneumonia is 4273, the normal class is 1989 and tuberculosis is 394. Then each image is done preprocessing with the image resizing so that the size is 32 × 32, then each RGB image is transformed into a grayscale image after that do the Gaussian standardization on each image by reducing the pixel value on the image with the mean of the entire image and divide it by the standard deviation of the entire image so that the mean of each image is approaching 0 and the standard deviation is approaching. The preprocessing process can be seen in Figure 2.

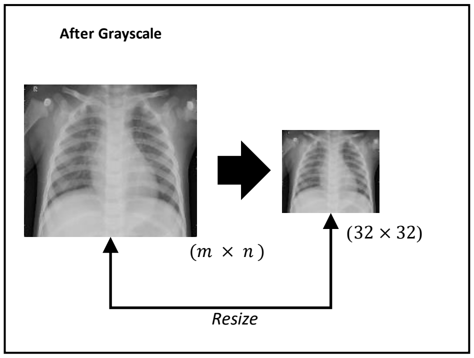

Figure 2. Preprocessing steps consist of the converting images to grayscale and resizing images from the original dataset.

After preprocessing step, we divide data into 80% as training and 20% as testing data. Then in the training data do sampling techniques, such as oversampling and undersampling to balance each class on the training data. Undersampling data is done by randomly throwing members of more data groups from fewer data groups, so the amount of data on each class is equally large, this undersampling process has shown in Figure 3 left side.

While oversampling data is the opposite of undersampling where fewer data groups of the data group are most likely to be reproduced randomly based on existing sample data so that the amount of data per group will be equally large, this oversampling process has shown on Figure 3 right side.

3. RESULTS AND DISCUSSIONS

We will evaluate the growth of loss value on the testing data to see when the model starts overfitting and when the model reaches the minimum value loss. This evaluates will be done from the start until the end of the training. The CNN models with a minimum loss value on testing data will be selected and then evaluated to see the performance of each model in predicting each data. After selecting the model by looking at the minimum loss value for each different sampling technique, the model performance will be seen based on the matrix classification and confusion matrix in the testing data.

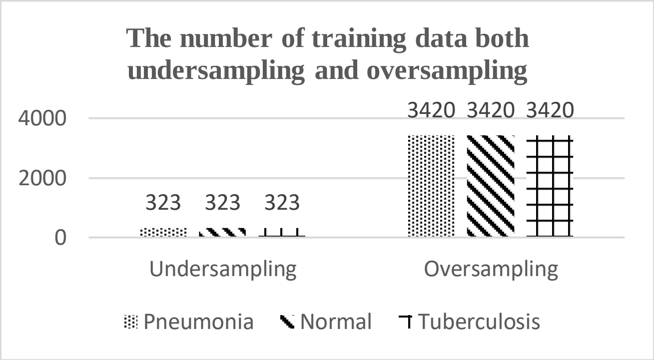

Figure 3. The number of training datasets after undersampling and oversampling.

3.1. Choosing The Best Model based on CNN Loss Model for Undersampling and Oversampling Datasets

Figure 4 shows that the performance of the model using the value of a loss on the training data continues to decline until the end of training data. The minimum loss value is 0.197. While in the testing data, the value of loss continues to decline until it reaches a minimum value on the 1781 epoch with a loss value of 0.373 and then converging afterward until the end of the training. So CNN's model is selected from the model of the undersampling of the training data that are generated in the 1781st epoch.

The changes of loss value on the training data and testing data per epoch can be seen in Figure 5. These loss values result from the CNN models that are formed using the training data from the oversampling results. From figure 5, we can see the model's performance based on the value of a loss on the training data. These values tend to decline until the 1999th epoch with a minimum loss value is 0.020 and converging afterward until the end of the training. While in the testing data, the value of loss tends to decline until it reaches a minimum on the 596th epoch with a value of the loss is 0.244 and then increases until the end of the training. So CNN's model from the oversampling with the selected training dataset is the model that is generated on the 596th epoch.

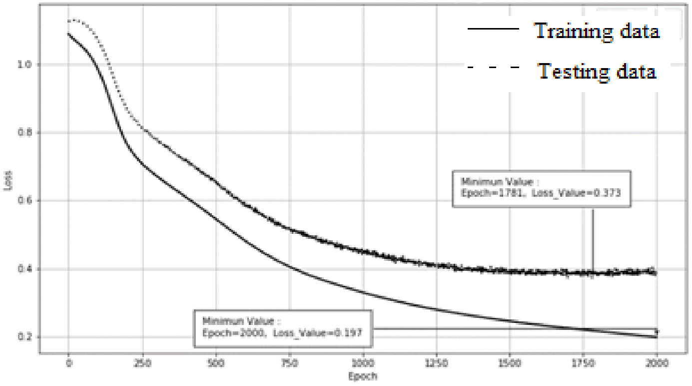

Figure 4. The value of the CNN loss model from undersampling.

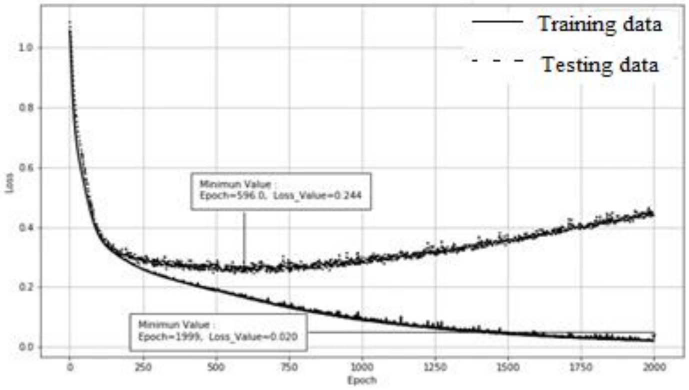

Figure 5. The value of the CNN loss model from oversampling.

3.2. The Accuracy during Training of Undersampling and Oversampling Dataset

Table 2 shows the matrix classification of the CNN Model for the undersampling dataset. In this table, CNN's model with undersampling dataset has accuracy for micro averages precision and a micro average recall is 0.86. While the macro average precision has a value of 0.75 and a macro average recall of 0.86. In the class of pneumonia has the highest precision of 0.94 followed by normal 0.80 then tuberculosis 0.49. Although tuberculosis has the lowest precision value compared to other classes, tuberculosis has the highest recall value of 0.96 followed by pneumonia 0.92 then normal 0.72.

Table 2. Matrix classification of the CNN Model for the undersampling dataset.

	Precision	Recall	Number of Data
Pneumonia	0.94	0.92	853
Normal	0.80	0.72	408
Tuberculosis	0.49	0.96	71
Accuracy	0.86		1332
Micro Average	0.86	0.86	1332
Macro Average	0.75	0.86	1332

The matrix classification of the CNN model for oversampling dataset is described in Table 3. This table shows that the CNN model with oversampling has accuracy, micro averages precision and a micro average recall of 0.90. While the macro average precision has a value of 0.80 and a macro average recall of 0.87. The Precision and the highest recall rate are in the pneumonia disease class with the values of 0.95 and 0.96. For the second-highest precision that is in the normal class with a value of 0.88, then tuberculosis with a precision value of 0.58. While the second-best recall value in tuberculosis class with 0.85 and normal value with recall value 0.80.

3.3. Model Evaluation of Undersampling and Oversampling dataset

Many problems in the study to detect a disease, more prioritizing the measurement in the case of recall [14]. In this study, recall is important because high recall means that the CNN Model made has a slight error rate in detecting a person who affected by pneumonia or tuberculosis.

The formation of the Receiver Operating Characteristics (ROC) curve to calculate the AUC value for each model is done to choose the best model. The ROC curve is commonly used in medical decision making and also used in the selection of the best models of machine learning algorithms and research in the field of data mining. The ROC curve summarizes and displays informally from the performance of a model especially recall (Figure 6 and 7).

Table 3. Matrix Classification of the CNN Model for oversampling dataset.

	Precision	Recall	Number of Data
Pneumonia	0.95	0.96	853
Normal	0.88	0.80	408
Tuberculosis	0.58	0.86	71
Accuracy	0.90		1332
Micro Average	0.90	0.90	1332
Macro Average	0.80	0.87	1332

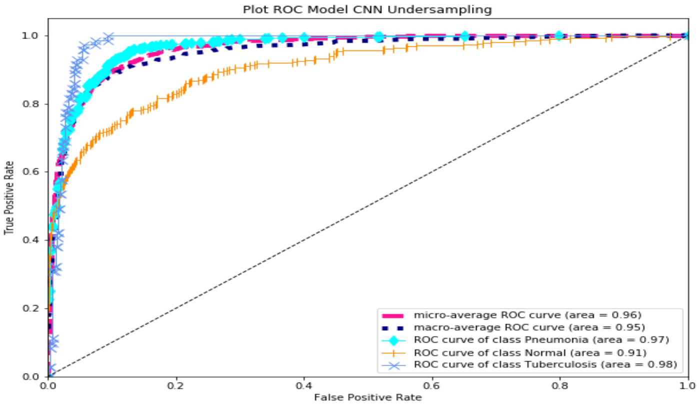

Figure 6. The result of ROC and AUC for the best CNN model on an undersampling dataset.

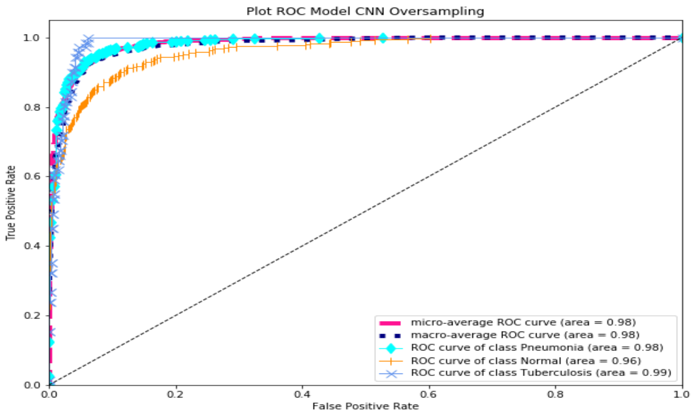

Figure 7. The result of ROC and AUC for the best CNN model on the oversampling dataset.

From Figure 6 and Figure 7, we can see the model's performance using the AUC value. It appears that the CNN model resulted from the oversampling with the training dataset is better than the CNN model that resulted in from undersampling with the training data. The value of micro average AUC, macro average AUC, AUC class of pneumonia, normal grade AUC and AUC of class tuberculosis on CNN oversampling model are 0.98, 0.98, 0.98, 0.96, and 0.99, respectively. While on CNN model undersampling the value of micro average AUC, the macro average AUC, AUC class of pneumonia, normal class AUC and AUC class tuberculosis are 0.96, 095, 0.97, 0.91, and 0.98, respectively. Since every AUC value on CNN's oversampling model is greater than the CNN undersampling model, the best model based on the sampling technique is a CNN model created by the oversampling with the training dataset.

4. CONCLUSIONS

In this paper, CNN is designed to diagnose lung diseases. The imbalanced dataset has been solved using undersampling and oversampling techniques on the training dataset. The results showed that the best of the CNN model is the model resulted from the oversampling, where the AUC of class tuberculosis is 0.99 and the AUC of pneumonia class is 0.98. The CNN model with oversampling can identify 86% of tuberculosis and 96% of pneumonia disease correctly on the testing dataset.

REFERENCES

[1] "Tuberculosis." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis. [Accessed: 08-Jul-2019].

[2] "Pneumonia." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia. [Accessed: 08-Jul-2019].

[3] C. Qin, D. Yao, Y. Shi, dan Z. Song, "Computer-Aided Detection in Chest Radiography Based on Artificial Intelligence: A Survey," *BioMedical Engineering Online*, vol. 17, no. 1, pp. 1–23, 2018. https://doi.org/ 10.1186/s12938

[4] A. M. Schilham, B. van Ginneken, dan M. Loog, "A Computer-Aided Diagnosis System for

Detection of Lung Nodules in Chest Radiographs with An Evaluation on A Public Database," *Medical Image Analysis*, vol. 10, no. 2, pp. 247–258, 2006.

[5] J. H. Tan, U. R. Acharya, C. Tan, K. T. Abraham, dan C. M. Lim, "Computer-Assisted Diagnosis of Tuberculosis: A First Order Statistical Approach to Chest Radiograph," *Journal of Medical Systems*, vol. 36, no. 5, pp. 2751–2759, Oktober. 2012.

[6] S. Hwang, H.-E. Kim, J. Jeong, dan H.-J. Kim, "A Novel Approach for Tuberculosis Screening Based on Deep Convolutional Neural Networks," *In: Medical imaging 2016: computer-aided diagnosis*, Vol. 9785, San Diego: SPIE; 2016. https://doi.org/10.1117/12.2216198.

[7] Y. Bar, I. Diamant, L. Wolf, dan H. Greenspan, "Deep Learning with Non-Medical Training Used for Chest Pathology Identification," *In: SPIE medical imaging; Orlando, Florida, United States. SPIE*; 2015. pp. 7. https://doi.org/ 10.1117/12.2083124.

[8] R. H. Abiyev dan M. K. S. Ma'aitah, "Deep Convolutional Neural Networks for Chest Diseases Detection," *Journal of Healthcare Engineering*, vol. 2018, Article ID 4168538, 2018. Available: https://doi.org/10.1155/2018/4168538. [Accessed: 31-Jul-2019].

[9] M.Paul, "Chest X-Ray Images (Pneumonia) | Kaggle." [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. [Accessed: 31-Jul-2019].

[10] M.Kevin, "TBC." [Online]. Available: https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities.

[11] D. S. Kermany, K. Zhang, dan M. H. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification," *Mendeley data 2*, 2018.

[12] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treafigu Diseases by Image-Based Deep Learning.," *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018.

[13] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, dan G. Thoma, "Two Public Chest X-Ray Datasets for Computer-Aided Screening of Pulmonary Diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–7, Dec. 2014.

[14] Y. Zhao, Z. S.-Y. Wong, dan K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," *Journal of Healthcare Engineering*, vol. 2018, p. 6275435, 2018.