IMPLEMENTATION OF CONVOLUTIONAL NEURAL NETWORK (CNN) ALGORITHMIC FOR COVID-19 DISEASE AND PNEUMONIA X-RAY IMAGE CLASSIFICATION

Fitria Agustina[1], Andi Farmadi[2], Dwi Kartini[3], Dodon Turianto Nugrahadi[4], Triando H. Saragih [5]
[12345]Ilmu Komputer FMIPA ULM
A. Yani St. KM 36 Banjarbaru, South Kalimantan
Email: j1f114203@mhs.ulm.ac.id[1], andifarmadi@ulm.ac.id[2],
dwikartini@ulm.ac.id[3], dodonturianto@ulm.ac.id[4],
triando.saragih@ulm.ac.id [5]

Abstrak

Pneumonia caused by the corona virus is different from ordinary pneumonia. One way to find out which pneumonia is caused by the corona virus is to do an X-ray. The disadvantage of this examination is that it requires a radiologist and the analysis time is relatively long. Therefore, to overcome this problem, deep learning methods can be used by implementing the Convolutional Neural Network (CNN) Algorithm method for X-ray image classification. The implementation of the Convolutional Neural Network (CNN) Algorithm is done by using training data of 4800 images which are trained using batch size values of 16, 32, and 64. The train process with batch size values of 16, 32 and 64 produces an average accuracy of 90%, 91% and 92%, while the loss values are 0.22, 0.16 and 0.25. From this process it was found that batch 64 was the best loss and accuracy result for training data. The test data with batch values of 16, 32, and 64 resulted in an accuracy of 76%, 82% and 76%, while the loss values were 0.79, 0.53 and 0.63. The results of this manual testing of 30 photos contained 7 images that are not recognized by the model because of the images look similar to each other with an accuracy of 76%. From this process it was found that batch 32 was the best loss and accuracy result for testing data.

Keywords : *Covid-19, Pneumonia, CNN, Machine Learning, Classification.*

1. INTRODUCTION

At the beginning of 2020, the world was shocked by an outbreak of a new pneumonia that started in Wuhan, Hubei Province, which then spread rapidly to more than 190 countries and territories. This outbreak was named Coronavirus disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-AoV-2)[3]. Pneumonia caused by the corona virus is different from ordinary pneumonia even though the symptoms of the disease tend to be similar and almost indistinguishable [2]. One way to find out whether you have been exposed to the corona virus or not is to do an x-ray or x-ray. X-ray examination or x-ray is a medical imaging technique that uses electromagnetic radiation to take pictures or photos of the inside of the body. This procedure is part of the investigation for a more accurate diagnosis. The drawback is that performing x-rays and analyzing x-ray images requires a radiologist and is time-consuming. Therefore the development of automated analysis systems is necessary to save the precious time of medical professionals by using machine learning technology.

During this pandemic, machine learning technology or commonly known as machine learning can analyze X-ray results in seconds. This system will show whether the patient has a high risk of getting pneumonia from the corona virus or not. Usually radiologists take hours to days to review the scan results and write a report on it. So this system will be very useful for doctors to be able to interpret the scan in a short time.

In order to reach all of that, a machine learning tool is needed that is able to study the job more deeply. Therefore, the latest learning method is currently being developed that is able to study a job in more depth, namely deep learning. Deep learning is a part of machine learning that allows computers to learn based on past experience and understand commands based on given concepts[1]. One algorithm that applies deep learning methods is a deep neural network or better known as a Convolutional Neural Network (CNN).

2. RESEARCH METHOD

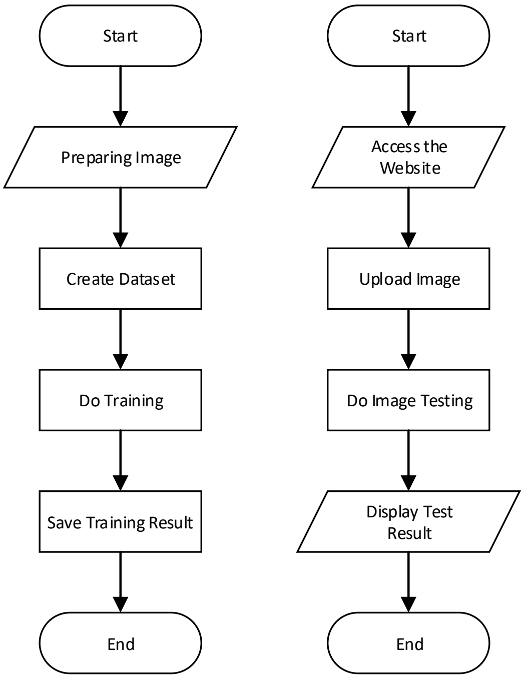


Figure 1. Research flow

The steps taken in this research are:

1. Determine the topic. At this stage, the search for theoretical references from books, journals, articles, research reports and websites on the internet is carried out for cases that will be used in research. The search research carried out includes, coronavirus and pneumonia, Machine Learning, Deep Learning, Convolutional Neural Network, image classification, multi-class classification and related literature as a reference in conducting this research.
2. Problem identification and formulation. at this stage, identification and formulation of the problem is one of the important elements to determine the quality of this research. The formulation in this study aims to determine the accuracy of the implementation of the Convolutional Neural Network algorithm in identifying x-ray images of covid and pneumonia diseases.
3. Determining research objectives tujuan
4. Determine the limitations and research methodology to avoid any deviations or widening of the subject matter so that the research is more focused and facilitates the discussion so that the research objectives will be achieved.
5. Conducting library studies related to the Convolutional Neural Network algorithm, deep learning, object recognition, covid, pneumonia.
6. Data collection of chest x-ray images of covid disease, chest x-ray images of pneumonia, and chest x-ray images of normal lungs taken from the google site page https://kaggle.com.
7. Preprocessing data by selecting what data will be used as input. A total of 6000 image data that have been labeled in each disease class will be used after the data selection process. Then the data will be divided into 2 with a ratio of 80%:20% to be used as train

data and test data. In each class, the train data will be 1600 and the test data will be 400. In addition, 100 disease image data for each class are provided to be used as data validation during the train process, so there are 300 data validations.

8. The CNN model design includes the number of layers used in the feature extraction layer and Fully Connected Layer, the number of filters, the size of the kernel, the use of Pooling and the activation function. In this research, the CNN model made consists of 1 convolution layer, 2 Pooling layers and 1 fully connected layer. The number of filters used in the convolution layer is 64 filters. The kernel size used in each convolution layer is 3x3 and in each Pooling layer is 2x2. The activation functions used are ReLU and softmax.
9. CNN model testing. In this stage, the model is tested using the saved model from the results of the model training process using the cnn.evaluate function in python. And also inputting images manually into the web application.
10. The accuracy results obtained will be based on the results of the train and test data.

3. RESULTS AND ANALYSIS

3.1 Data Collecting

The data used in this study are labeled chest x-ray images and obtained by downloading on the internet page https://kaggle.com. After downloading the data obtained quite a lot with a total of 21,700 labeled chest x-ray image data. Each disease class has a different amount of data, namely: 4,307 images for the COVID19 disease class, 5,628 images for the Pneumonia disease class, and 11,775 images for the normal lung class. This image data will be continued to the data selection stage so that it can be used as input to the CNN model that will be designed.

3.2 Preprocessing and Data Sharing

At this stage, manual selection of data is carried out by taking into account the provisions of the x-ray images that are included in certain disease classes. At this stage, after determining the terms and conditions of an x-ray image that can be used to detect covid and pneumonia, namely:

1. A covid x-ray image requires an x-ray image of the lungs that have a blurry rash on both lungs, this indicates that the lungs that are affected by the covid disease cause damage to both of the patient's lungs.
2. The x-ray image needed for pneumonia is an x-ray image with a rash on one of the lungs, this indicates that the lung with pneumonia is marked by damage to one of the patients.
3. X-ray images of normal lungs with features on both lungs in solid black without any blurring of white.

After selecting data based on the above criteria for each class from the 21,700 x-ray images obtained, the researcher selected 2000 images for each disease class manually and sorted one by one according to the criteria for x-ray images in each disease to be used as input for training data and test data.

Furthermore, after as many as 6000 x-ray images have been selected, data balancing is carried out to balance the data that will be entered in the model input. The classification usually relies heavily on training data. The training data itself, the number of data distributions for each class very rarely has the same amount. In real conditions, it is very often encountered that the number of datasets for each class is different. This condition is known as unbalanced data or data imbalance.

3.3 CNN Model Design

In this study, the initial parameters used in this research design are, 1 convolutional layer, filters with a value of 64, the padding value is 'same', the number of kernels used is 3, the activation used is ReLu and softmax, the channel used is 3, There are 2 pooling layers used with a shift of 2, flatten, and also dense with a value of 3. The parameter values can be seen in the table below.

Table 1. Parameter

Parameter	Value
Convolution Layer	1
Fillters	64
Padding	Same
Kernel	3
Pooling Layer	2
Chanel	3
Dense	3
Flatten	1
Aktivation ReLu	1
Aktivation Softmax	1

After importing all the required libraries, the Convolution2D sublibrary is used to start the CNN in the first stage of the convolution process. Because we process data in the form of images, we use a special convolution library for 2 dimensions, namely Convolution2D. Then the MaxPooling2D Sublibrary is used after the convolution process, by taking the maximum value (called maxpooling). Because we are processing images, we use maxpooling for 2 dimensions. The Flatten sublibrary is used for the flattening process that is carried out after the maxpooling process. the next step is to "flatten" or reshape the feature map into a vector so that it can be used as input from the fully-connected layer. Dense sublibrary is used to define the parameters of neural networks. Dense is a funsdi to add a fully connected layer.

3.4 CNN Model Train

After loading the CNN model, the next step is to set the hyperparameters. Hyperparameters are parameters that are set before being used in the training process, the purpose of this setting will affect how well the model will be trained. The hyperparameters used can be seen in the following table:

Table 2. Hyperparameter

hyperparameter	value
Number of epoch	100
Number batch size	16, 32, 64

Based on the table above, it can be explained that in this study, hyperparameter settings were carried out on the number of epochs and batch size. The epoch is a hyperparameter that determines the frequency with which the learning algorithm will work across the training data set. Batch size determines the number of samples to be distributed over the network. After setting the hyperparameters, the next step is to train the model using the scenario of setting the hyperparameter values that have been set in the table using each training data that will be entered into the model and trained.

3.5 Result from CNN Model Train

By using epochs of 100 and the number of batch sizes as many as 16, 32, and 64 and then getting the following accuracy results.

1. *Batch size 16*

   The results of the training data used a batch size of 16 with the time required to train 4800 images for 28 hours 31 minutes 47 seconds. The results on training data with batch size 16 are the average results of the accuracy of the train results with batch 16 values of 90% and loss of 0.22. Comparison of accuracy and loss values for training data with batch size 16 values can be seen in Figure 2.

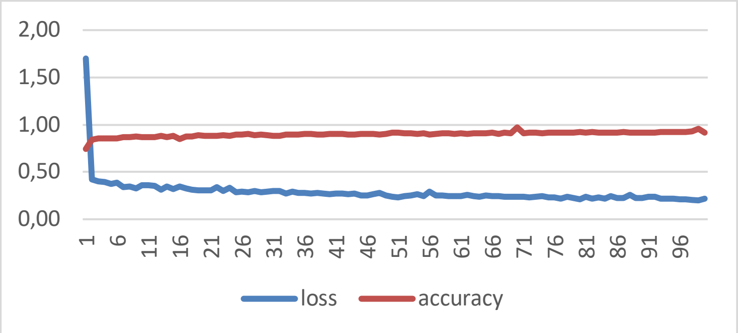


Figure 2. comparison of accuracy value and loss value of batch size 16

It can be seen from the graph, that the accuracy of the training results is stable and increases little by little, the graph also shows that the best accuracy value is obtained at the 70th epoch with an accuracy value of 97%. In addition, for the loss value from epoch 1 to wpoch 100, the graph decreases until it reaches a loss value of 0.22.

2. *Batch size 32*

The results of the training data used a batch size of 32 with the time required to train 4800 images for 26 hours 6 minutes 26 seconds. it can be seen that what is obtained in batch training data is the average result of the accuracy of the train results with a batch value of 32 of 91% and a loss of 0.16. Comparison of accuracy and loss values for training data with batch size 32 values can be seen in Figure 3.

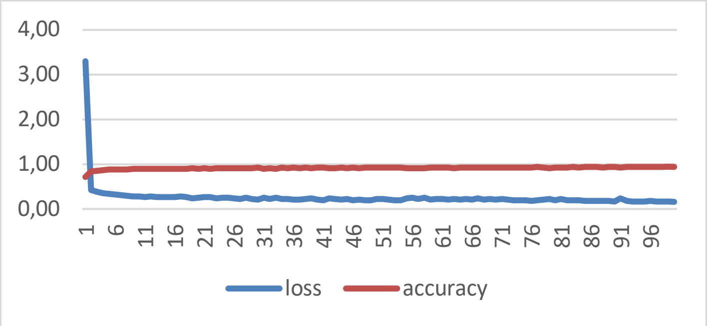


Figure 3. comparison of accuracy value and loss value of batch size 32

It can be seen from the graph, that the accuracy of the training results is stable and increases little by little, the graph also shows that the best accuracy value is obtained at the 99th epoch with an accuracy value of 94%. In addition, for the loss value from epoch 1 to wpoch 100, the graph decreases until it reaches a loss value of 0.16.

3. *Batch size 64*

The results of the training data use a batch size 64 value with the time needed to train 4800 for 25 hours 28 minutes 22 seconds. it can be seen that what is obtained from the training data with batches is the average result of the accuracy of the train results with a batch value of 63 of 92% and a loss of 0.25. Comparison of accuracy and loss values for training data with batch size 64 values can be seen in Figure 4.

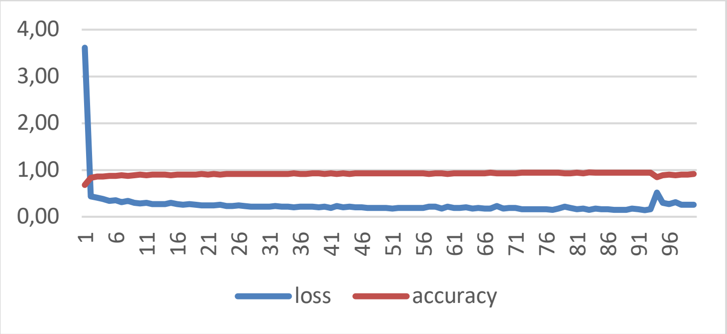


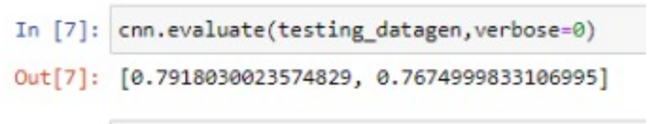
Figure 4. comparison of accuracy value and loss value of batch size 64

It can be seen from the graph, that the accuracy of the training results is stable and increases little by little, the graph also shows that the best accuracy value is obtained at the 92nd epoch with an accuracy value of 95%. In addition, for the loss value from epoch 1 to wpoch 100, the graph decreases until it reaches a loss value of 0.25.

3.6 CNN Model Testing

After training the model, the next step is to test the model using the cnn.evaluate function in python. Tests were carried out on each test data from 2 scenarios, namely data testing with each test 3 times using different batch size values. The batch size values used are 16, 32, and 64. With the number of epochs of 100. Then the next test is carried out manually using direct input on the web application using 10 image data in each class, so that the total images in this test are 30 image data.

In the test using the first scenario, the test data uses a batch size of 16, the accuracy results obtained are 76% with a data loss of 79%.

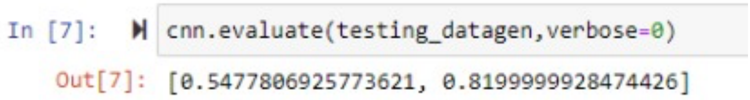


```
In [7]:  cnn.evaluate(testing_datagen,verbose=0)

Out[7]:  [0.7918030023574829, 0.7674999833106995]
```

Figure 5. test data results with batch size 16

Then in testing the test data using a batch size value of 32, the accuracy results obtained are 76% with a data loss of 79%.

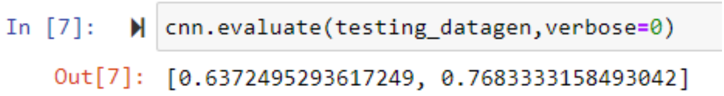


```
In [7]:     cnn.evaluate(testing_datagen,verbose=0)

Out[7]:  [0.5477806925773621, 0.8199999928474426]
```

Figure 6. test data results with batch size 32

Finally, in testing the test data using a batch size of 64, the accuracy results obtained are 54% with a data loss of 82%.



```
In [7]:     cnn.evaluate(testing_datagen,verbose=0)

Out[7]:  [0.6372495293617249, 0.7683333158493042]
```

Figure 7. test data results with batch size 64

Then the second test is carried out manually by entering several images that have never been used for training or testing and seeing the results of the model whether it can distinguish according to the training data. The amount of data used for manual testing is 10 image data in each disease class with a total of 30 images.

Based on the data from the tests carried out manually on the chest x-ray identification application for Covid and pneumonia diseases using the Convolutional Neural Network algorithm, 23 recognizable image data and 7 unrecognizable image data can be obtained. For the value of accuracy with an average of 78%. the accuracy value is obtained from the equation:

$$percentage\ accuracy = \frac{correct\ amount\ of\ test\ data}{total\ number\ of\ test\ data} \times 100\%$$

$$= \frac{23}{30} \times 100\%$$

$$= 76\%$$

From the calculation above, it can be seen that the accuracy rate of the Convolution Neural Network method in identifying chest x-rays for COVID and pneumonia is 78%.

4. CONCLUSION

From the results of research that has been carried out with model testing, it can be concluded that the results of this study using training data as many as 4800 images trained using batch size 16, 32, and 64 values, namely, with a batch value of 16 the average accuracy produced is 90 % and the loss value is 0.22. Of the 100 trained epochs, in batch 16 the best accuracy was obtained at the 70th epoch with an accuracy of 97%. The results of the train data using a batch value of 32 produce an average accuracy of 91% and a loss value of 0.16. Of the 100 trained epochs, at the batch value of 32, the best accuracy was obtained at the 99th epoch with an accuracy of 94%. Finally, the results of the train data using a batch value of 64 resulted in an average accuracy of 92% and a loss value of 0.25. Of the 100 trained epochs, at the batch value of 64, the best accuracy was obtained at the 92nd epoch with an accuracy of 95%.

Furthermore, for the accuracy of the test data for the batch size 16 value, the accuracy obtained is 76% and the loss value is 0.79, for the batch size 32 value the accuracy is 82% and the loss value is 0.54, for the batch size 64 value the accuracy is obtained by 76% and a loss value of 0.63. The results of this manual testing of 30 photos there are 7 images that are not recognized by the model because the images look similar to each other with an accuracy of 76%.

REFERENCES

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning Series)*. The IMT Press.

[2] Relman,E.(2020)*.BusinessinsiderSingapore.CitedJan28th2020.Available on:https://www.businessinsider.sg/deadly-china-wuhan-virusspreading-human-to-human-officials-confirm-2020-1/?r=US&IR=T.*

[3] Susilo, dkk. (2020). *Coronavirus Disease 2019: Tinjauan Literatur Terkini*. jakarta:Departemene ilmu penyakit dalam fakultas kedokteran universitas indonesia.