

# Performance Comparison of Deep Learning Models to Detect Covid-19 Based on X-Ray Images

S Riyadi\*<sup>1</sup>, Y Lestari<sup>2</sup>, C Damarjati<sup>3</sup>, K H Ghazali<sup>4</sup>

<sup>1,2,3</sup>Department of Information Technology, Universitas Muhammadiyah Yogyakarta, Indonesia.

<sup>4</sup>Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang, Malaysia.

E-mail: riyadi@umy.ac.id, yunitaysz123@gmail.com, cahyadamarjati@umy.ac.id, kamarul@ump.edu.my

**Submitted: 14 Jan 2022, revised: 7 Feb 2022, accepted: 24 Feb 2022**

**Abstract** The SARS-Cov-2 outbreak caused by a coronavirus infection shocked dozens of countries. This disease has spread rapidly and become a serious threat. It even destroys various sectors of life. Along with technological developments, various deep learning models have been developed to classify between Normal and Covid-19 from chest X-ray images, such as Inception V3, Inception V4, and MobileNet. These models have been separately reported to perform good classification on Covid-19. However, there is no comparison of their performance in classifying Covid-19 on the same data. This research aims to compare the performance of the three mentioned deep learning models in classifying Covid-19 based on X-ray images. The methods involve data collection, pre-processing, training, and testing using the three models. From 2,169 datasets, the three models can classify the Covid-19 based on X-ray images. The result showed that the MobileNet model achieved the best performance with an average accuracy of 99.67%, precision of 99.77%, recall 99.38%, specificity 99.38%, and f-score 99.67%. For the Inception V3 model, it was obtained an average accuracy of 99.62%, the precision of 99.65%, recall value 99.5%, specificity 99.5%, and f-score 99.52%; while the InceptionV4 model obtained an average accuracy of 97.79%, the precision of 98.11%, recall of 90.18%, specificity of 90.18%, and f-score of 97.25%. Furthermore, the MobileNet model showed stable performance in achieving graphic results since it had extensive layers. The more layers the model has, the better the accuracy is obtained.

**Keywords:** Covid-19, Deep Learning, InceptionV3, InceptionV4, MobileNet

## 1. Introduction

Covid-19 (Coronavirus Disease 2019) is a group of pneumonia cases with unknown causes. It was firstly identified in Hubei province, China, in December 2019. The new type of coronavirus has spread rapidly to become a new pandemic. The disease symptoms are usually fever, cough, shortness

of breath, and fatigue [1], [2], [3]. The spread of the Covid-19 disease is a severe threat to countries globally that can disrupt and even destroy various sectors of life [4], [5].

Along with technological developments, an examination for Covid-19 detection has utilized imaging modalities such as X-ray, CT-Scan images, and an artificial neural network [6], [7]. One type of artificial neural network used in the Covid-19 detection is deep learning, and the most popular deep learning algorithm is Convolution Neural Network (CNN) [8], [9], [10], [11]. Deep learning creates a model of Covid-19 and Normal X-ray images, then uses it to classify the X-ray images. So, the application of deep learning helps detect the Covid-19 disease early on [12].

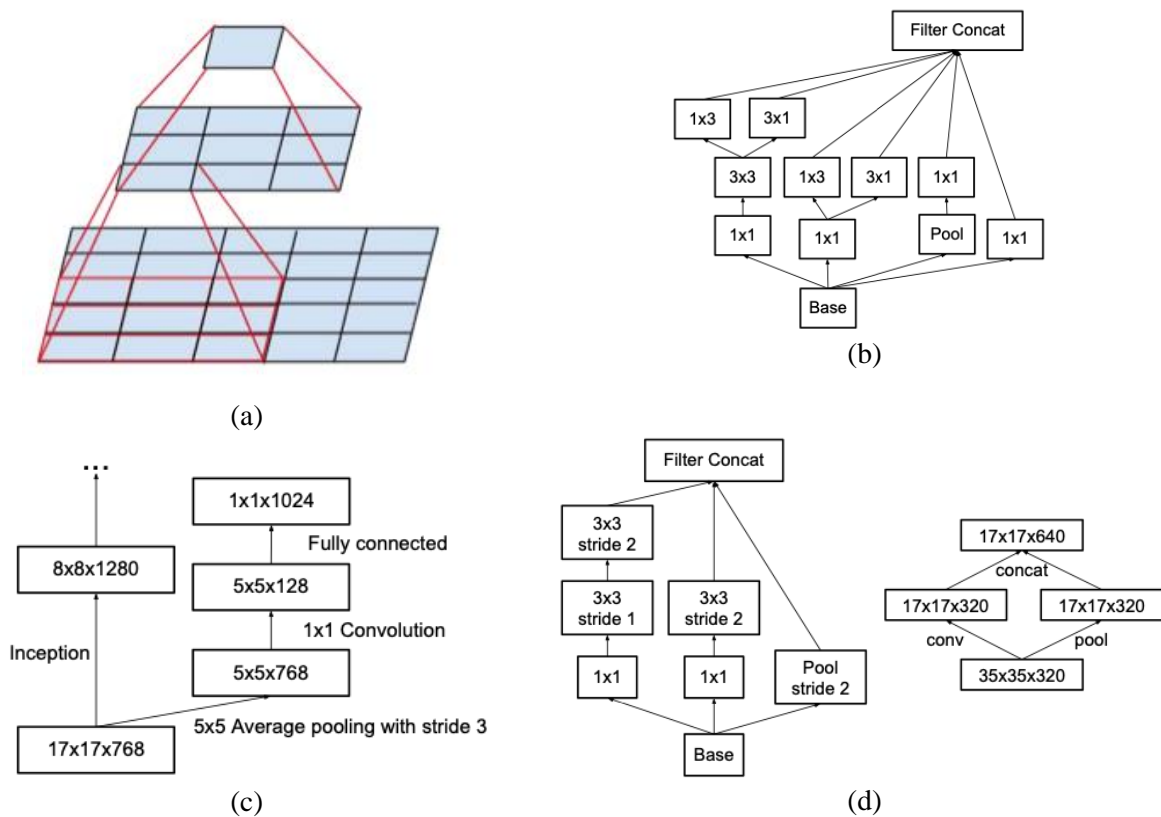
Researchers have initiated the analysis and the detection of Covid-19 using deep learning X-Ray images since 2020. The researchers used Convolutional Neural Network models including VGG16, VGG19, InceptionV3, MobileNetV2, ResNet50, and DenseNet121 to classify Normal, Pneumonia, and Covid-19 chest X-ray images [6]. They classified the images into two groups, namely Covid-19 or Normal, Covid-19 or Pneumonia, and Normal or Pneumonia, using thousands of publicly available chest X-ray images. Among the models, Inception V3 and MobileNet performed good classification results. Another research group proposed a pre-trained model CNN using ResNet50, InceptionV3, and Inception\_ResNetV2 [13]. Since the dataset was only 50 Covid-19 and 50 Normal X-ray images, they included transfer learning techniques to overcome the limited dataset. Further research on Covid-19 used the CNN pre-trained model VGG16 and InceptionV3. This research intended to develop a fast, accurate, and low-cost diagnostic system to detect Covid-19 using chest X-Ray. The dataset consisted of 2,905 chest X-ray images with 219 confirmed positive cases of Covid-19, 1345 positive pneumonia, and 1,361 normal images. The result showed that the Inception V3 model provided the highest accuracy of 99.35% for the two binary classifications (Normal vs Covid-19 and Covid-19 vs pneumonia) compared to the VGG16 model accuracy of 97.71% [14].

The Inception V3 and MobileNet are two deep learning models which perform good classification results among the deep learning models [6], [13], [14]. In the advancement of Inception V3, the Inception V4 simplifies the Inception V3 with a more uniform architecture [15]. However, a performance comparison between them on the same dataset to classify Covid-19 was not reported. Therefore, this study aims to compare their performance in terms of accuracy, precision, recall, specificity, and F-score, using the same X-ray images dataset.

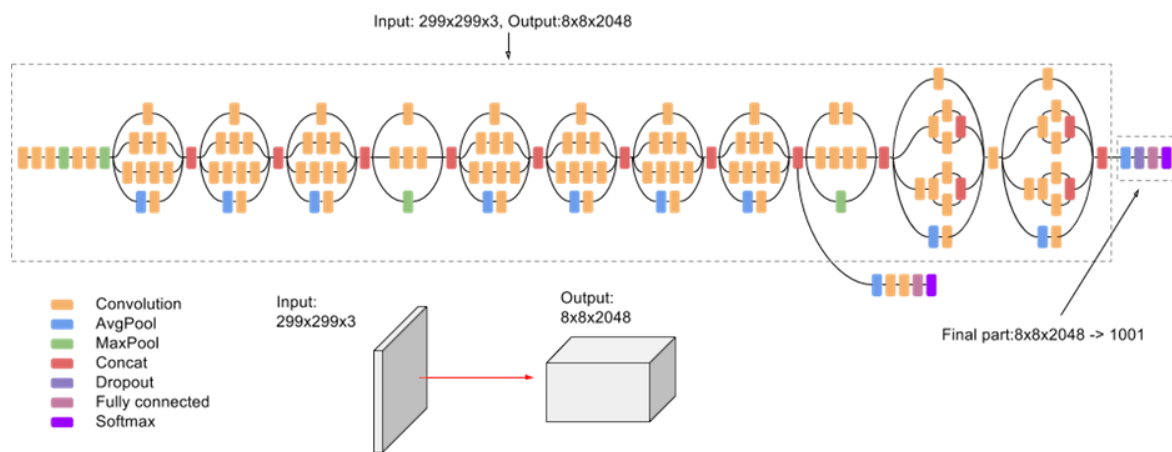
## 2. Deep Learning Framework

Deep Learning is an artificial intelligence that utilizes artificial neural networks to learn the characteristics of large datasets and provides a very robust architecture for supervised learning. In machine learning, there are techniques for using feature extraction from training data and unique learning algorithms to classify images and recognize sounds [16]. Deep learning processing requires datasets and larger datasets can give more accurate results.

InceptionV3 is a convolutional neural network architecture to analyze an image and detect an object. Compared to the previous version, Inception V3 focuses on increasing computational efficiency on the number of parameters and hardware resources cost. Figure 1 shows the Inception V3 architecture including the step-by-step parts: (1) factorized convolution to increase computational efficiency, (2) smaller scale convolutions to train faster, (3) asymmetric convolutions to reduce the number of parameters, (4) an auxiliary classifier which is a small CNN regulator inserted between layers, and (5) grid resize reduction [17], [18]. The complete architecture is shown in Figure 2 [17].



**Figure 1.** Step-by-step development of Inception V3: (a) smaller scale convolutions for faster training, (b) asymmetric convolutions to reduce the number of parameters, (c) an auxiliary classifier which is a small CNN regulator inserted between layers, and (d) grid resize reduction [17], [18].

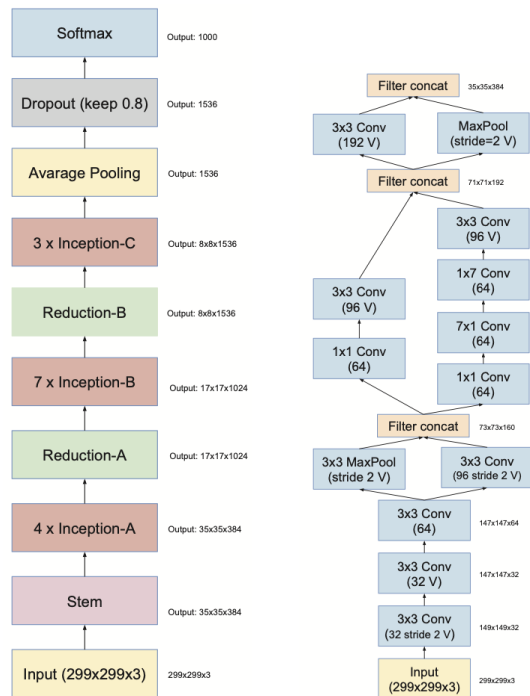


**Figure 2.** The complete architecture of Inception V3 [17]

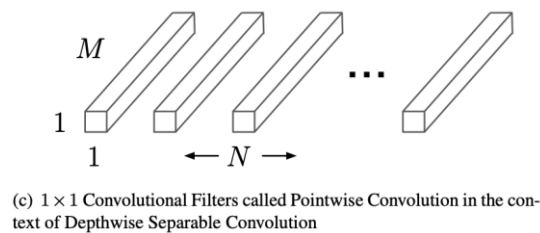
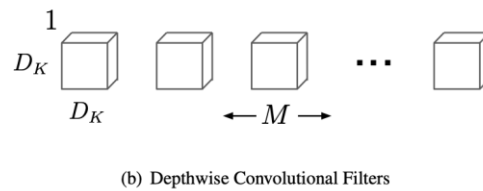
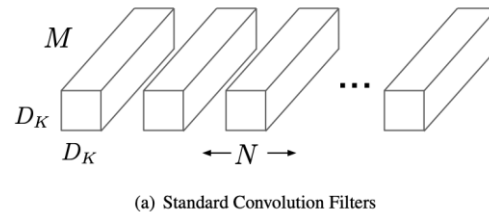
InceptionV4 is a pure artificial neural network architecture with no residual connections with approximately the same image recognition performance as Inception-ResNet-v2. Inception V4 has a simple architecture with more uniform modules than Inception V3 as shown in Figure 3 [19].

Inception V4 has three modules, namely Inception-A, Inception-B, and Inception-C, which continues by average pooling and ends with a fully connected layer as a classification layer.

MobileNet is a deep learning architecture that can address a large amount of data. The essential MobileNet architecture is the use of a particular layer called depth-wise separable convolution, which reduces complexity and reduces parameters to produce a larger model, as shown in Figure 4 [20].



**Figure 3.** The overall scheme of Inception V4 [19]



**Figure 4.** (a) Standard convolution filter is replaced by (b) depth-wise and (c) pointwise convolution [20]

### 3. Method

This research involved five main stages: data collection, pre-processing, training, testing, and result analysis. The data used in this study were X-Ray image data obtained from kaggle.com. The data obtained amounted to 2,169 X-Ray images that were divided into two types, namely Covid-19 and Normal. Pre-processing is a stage of image preparation to make the data appropriate for the training and testing. It included image resizing to 224x224 pixels and folding images into 5-fold datasets for training and testing. A k-fold is a method to evaluate the model performance or algorithm [21].

The training stage created a deep learning model of the Covid-19 classifier. The configuration included a stochastic gradient descent optimizer with LR as much as  $1e-4$ , momentum as much as 0.9, and loss using categorical\_crossentropy. The training process also used an epoch of 100 and a batch size of 32. The model was then saved and used for a validation process. When the training was conducted, the validation was carried out based on 5-fold using a 20% dataset to confirm whether the training models were good or not.

The analysis in this study compared the performance of the training, validation and testing based on three models using the confusion matrix. The confusion matrix was calculated based on two classes, namely Covid-19 and Normal, shown in **Table 1**, where the comparison of the performance matrix included the parameters of accuracy, precision, recall, specificity, and f-score following Table

2. The parameters used in the analysis were equalizing all classes in each fold. The higher the value obtained, the better the model performed. For the training, the epoch value of graphic images and accuracy were compared. Meanwhile, for the testing, the results of all performance parameters were compared. The higher the value of performance matrix, the better and more efficient the performance generated by the model.

**Table 1.** Confusion matrix of Covid-19 and Normal class

	<b>Actual Covid-19</b>	<b>Actual Normal</b>
Predicted Covid-19	True Positive (TP)	False Negative (FN)
Predicted Normal	False Positive (FP)	True Negative (TN)

**Table 2.** Performance metrics for classification task [22]

<b>Metrics</b>	<b>Formula</b>	<b>Description</b>
Accuracy	$(TP+TN)/(TP+FP+TN+FN)$	Measures the ratio of correct prediction over the total data
Precision	$TP/(TP+FP)$	Measures the positive pattern that are correctly predicted from the positive class
Recall	$TP/(TP+TN)$	Measures the fraction of positive patterns that are correctly classified
Specificity	$TN/(TN+FP)$	Measures the fraction of negative patterns that are correctly classified
F-score	$2*Precision*Recall/(Precision+Recall)$	Measures the harmonic mean between recall dan precision

#### 4. Result and Discussion

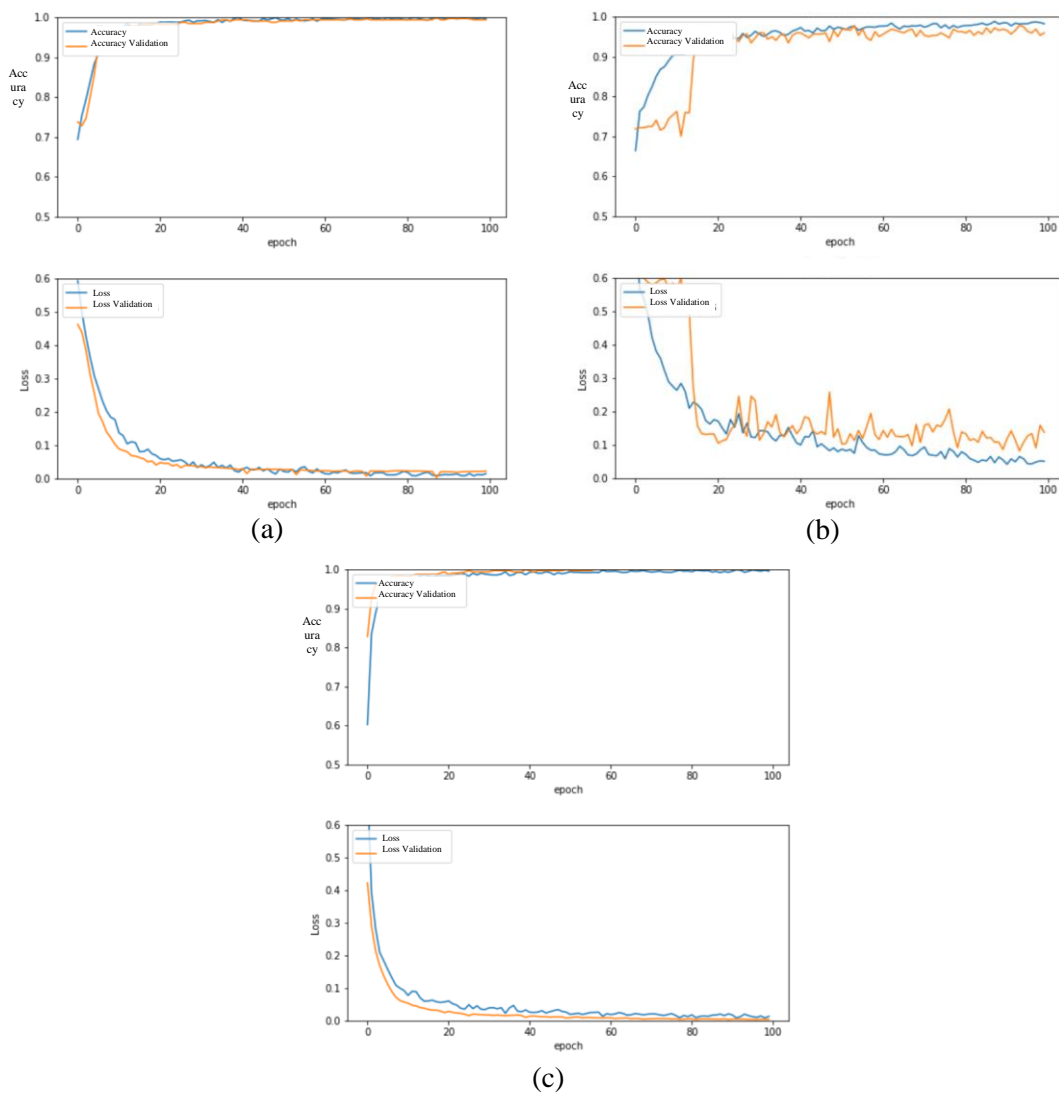
The dataset was divided into 1,388 images for training, 347 images for validation, and 434 for testing. All dataset was used for training, validation and testing of each model. The validation results on 347 images using each model are shown in Table 3. The validation of the training accuracy resulted in 99.68% for Inception V3 and MobileNet, whereas the value slightly decreased to 97.93% for Inception V4. For the training loss, the validation resulted in 0.013%, 0.072%, and 0.012% for Inception V3, Inception V4, and MobileNet respectively. According to the accuracy and loss metric in validation process, Inception V3 and MobileNet have better training results than Inception V4. It may be affected by the simpler architecture of Inception V4 than Inception V3 and MobileNet.

**Table 3.** Validation result of each model

Fold	Accuracy			Loss		
	Inception V3	Inception V4	MobileNet	Inception V3	Inception V4	MobileNet
1	99.77%	97.24%	99.31%	0.007%	0.144%	0.023%
2	99.77%	97.47%	99.54%	0.007%	0.092%	0.023%
3	99.77%	98.39%	99.54%	0.016%	0.037%	0.008%

4	99.54%	97.93%	100%	0.012%	0.038%	0.005%
5	99.54%	98.61%	100%	0.021%	0.050%	0.001%
<b>Average</b>	<b>99.68%</b>	<b>97.93%</b>	<b>99.68%</b>	<b>0.013%</b>	<b>0.072%</b>	<b>0.012%</b>

Figure 5 shows the validation graphs for training accuracy and loss for Inception V3, Inception V4, and MobileNet models. Inception V3 and MobileNet performed excellent and stable validation results on training accuracy and loss, while Inception V4 showed overshoot along with the epoch. Figure 5 is in accordance with the numeric result in Table 3. Since the Inception V4 has simpler architecture than Inception V3, the Inception V4 has small scale convolution and reduced grid size to perform faster training. In addition, the auxiliary classifier between layers in Inception V4 produce more stable training result.



**Figure 5.** Validation for training accuracy and loss of each model (a) Inception V3, (b) Inception V4, and (c) MobileNet

Table 4 presents the performance matrix of InceptionV3 testing on 343 images. The results showed that the largest value of the performance parameters was at fold = 3 (Precision and Specificity 100%) while the smallest was at fold = 5 (Precision, Recall and F-Score 99.20%). The results of the InceptionV3 calculation showed the average accuracy of 99.62%, precision of 98.94%, recall of 99.69%, specificity of 99.62%, and f-score of 99.30%. According to the testing result, the Inception V3 showed an excellent ability to classify Covid-19 and Normal X-ray images. Compared to Inception V3, Inception V4 performed slightly lower results, as shown in

**Table 5.** It resulted accuracy of 97.79%, precision 98.76%, recall 98.36%, specificity 98.17%, and F-score 95.92%. The superior results were obtained from the MobileNet model, as shown in Table 6. This model showed almost perfect ability indicated by the parameters' average with 99.67% accuracy, 98.76% precision, 100% recall, 99.55% specificity, and 99.37% F-score. The five parameters consistently agree with the superior ability of MobileNet to classify the data. This mostly caused by the depth-wise separable convolution which effectively handles a large amount of data training.

**Table 4.** Performance matrix for InceptionV3 testing

Fold	Accuracy	Precision	Recall	Specificity	F-Score
1	99.53%	98.30%	100%	99.37%	97.14%
2	99.76%	98.94%	100%	99.70%	99.46%
3	99.76%	100%	99.25%	100%	99.62%
4	99.53%	98.26%	100%	99.37%	97.12%
5	99.53%	99.20%	99.20%	99.67%	99.20%
Average	99.62%	98.94%	99.69%	99.62%	99.30%

**Table 5.** Performance matrix for InceptionV4 testing

Fold	Accuracy	Precision	Recall	Specificity	F-Score
1	97.23%	91.52%	98.18%	98.91%	94.73%
2	97.46%	90.52%	97.72%	97.39%	93.98%
3	98.38%	96.99%	97.72%	98.67%	97.35%
4	97.30%	93.91%	98.18%	97.83%	96.00%
5	98.61%	95.2%	100%	98.08%	97.54%
Average	97.79%	98.76%	98.36%	98.17%	95.92%

**Table 6.** Performance matrix for MobileNet testing

Fold	Accuracy	Precision	Recall	Specificity	F-Score
1	99.30%	97.45%	100%	99.05%	98.71%
2	99.53%	97.89%	100%	99.41%	98.93%
3	99.53%	98.49%	100%	99.33%	99.24%
4	100%	100%	100%	100%	100%
5	100%	100%	100%	100%	100%

---

Average	99.67%	98.76%	100%	99.55%	99.37%
---------	--------	--------	------	--------	--------

---

The ability of each model to classify the Covid-19 and Normal datasets is shown in Table 7. The table showed that Inception V3 and MobileNet had higher performance in classifying the Covid-19 and Normal X-ray images datasets. MobileNet resulted superior performance in accuracy, precision, and F-score, while Inception V4 performed slightly under Inception V3 and MobileNet models. This performance difference was because the MobileNet model had 87 layers, where the more the number of layers, the better the accuracy results. In addition, the MobileNet had a particular layer called Depthwise Separable Convolution, which was used to reduce the complexity. Therefore, the MobileNet model was more efficient and a very light model when compared to the InceptionV3 and InceptionV4 models in the Covid-19 classification.

**Table 7.** Confusion matrix of each model

<b>Performance Matrix</b>	<b>Class</b>	<b>Test</b>		
		<b>InceptionV3</b>	<b>InceptionV4</b>	<b>MobileNet</b>
<i>Accuracy</i>	Covid-19	99.62%	97.79%	99.67%
	Normal	99.62%	97.79%	99.67%
	<b>Mean</b>	<b>99.62%</b>	<b>97.79%</b>	<b>99.67%</b>
<i>Precision</i>	Covid-19	99.69%	98.36%	100%
	Normal	99.62%	97.87%	99.55%
	<b>Mean</b>	<b>99.65%</b>	<b>98.11%</b>	<b>99.77%</b>
<i>Recall</i>	Covid-19	98.14%	93.62%	98.76%
	Normal	99.86%	86.75%	100%
	<b>Mean</b>	<b>99.5%</b>	<b>90.18%</b>	<b>99.38%</b>
<i>Specificity</i>	Covid-19	99.86%	86.75%	100%
	Normal	98.14%	93.62%	98.76%
	<b>Mean</b>	<b>99.5%</b>	<b>90.18%</b>	<b>99.38%</b>
<i>F-Score</i>	Covid-19	99.31%	95.92%	99.57%
	Normal	99.73%	98.59%	99.77%
	<b>Mean</b>	<b>99.52%</b>	<b>97.25%</b>	<b>99.67%</b>

## 5. Conclusion

Three deep learning models, namely InceptionV3, InceptionV4 and MobileNet, have been implemented to classify Covid-19 X-ray images. The models performed excellent ability to classify Covid-19 and Normal X-ray images. The MobileNet model showed the most stable graph in the training phase compared to the InceptionV3 and InceptionV4 models. For the testing phase, MobileNet also had the highest accuracy (99.67%) compared to InceptionV3 (99.62%) and InceptionV4 (97.79%). Other performance criteria were also computed and revealed that the MobileNet was superior to Inception V3 and Inception V4 in classifying X-ray images. Researchers could conduct further research to compare with other deep learning models.



## 6. Acknowledgement

The authors acknowledge the Universitas Muhammadiyah Yogyakarta for the financial support of this research.

## References

- [1] O. L. Aiyegbusi, S. E. Hughes, G. Turner, S. C. Rivera, C. McMullan, J. S. Chandan, S. Haroon, G. Price, E. H. Davies, K. Nirantharakumar, E. Sapey, M. J. Calvert, "Symptoms, complications and management of long COVID: a review," *Journal of the Royal Society of Medicine*, vol. 114, no. 9, pp. 428–442, Jul. 2021, doi: 10.1177/01410768211032850.
- [2] M. Cascella, M. Rajnik, A. Aleem, et al. "Features, Evaluation, and Treatment of Coronavirus (COVID-19) [Updated 2022 Jan 5]". In: *StatPearls* [Internet]. Treasure Island (FL): StatPearls Publishing; Jan 2022. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK554776/>
- [3] J Elliott, M. Whitaker, B. Bodinier, O. Eales, S. Riley, et al. "Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people". *PLOS Medicine* 18(9): e1003777, 2021. <https://doi.org/10.1371/journal.pmed.1003777>
- [4] T. Laing, "The economic impact of the Coronavirus 2019 (Covid-2019): Implications for the mining industry," *The Extractive Industries and Society*, vol. 7, no. 2, pp. 580–582, 2020, doi: <https://doi.org/10.1016/j.exis.2020.04.003>.
- [5] B. Celik, S. Dane, and K. Ozden, "The Effects of COVID-19 Pandemic Outbreak on the Household Economy," *Journal of Research in Medical and Dental Science*, vol. 8, no. 4, pp. 51–56, Jul. 2020.
- [6] M. M. Ahsan, R. Nazim, Z. Siddique, & P. Huebner. Detection of COVID-19 Patients from CT Scan and Chest X-ray Data Using Modified MobileNetV2 and LIME. *Healthcare (Basel, Switzerland)*, 9(9), 1099, 2021. <https://doi.org/10.3390/healthcare9091099>
- [7] B. Sekeroglu and I. Ozsahin, "Detection of COVID-19 from Chest X-Ray Images Using Convolutional Neural Networks," *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 25, no. 6, pp. 553–565, Sep. 2020, doi: 10.1177/2472630320958376.
- [8] A.H. Reshi, F. Rustam, A. Mehmood, A. Alhossan, Z. Alrabiah, A. Ahmad, H. Alsuwailem, G. S. Choi, "An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification", *Complexity*, vol. 2021, Article ID 6621607, 12 pages, 2021. <https://doi.org/10.1155/2021/6621607>
- [9] Thakur S, Kumar A. X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN). *Biomed Signal Process Control*. 2021 Aug;69:102920. doi: 10.1016/j.bspc.2021.102920. Epub 2021 Jun 30. PMID: 34226832; PMCID: PMC8241644.
- [10] S. J. S. Rajasekar, V. Narayanan, V. Perumal. Detection of COVID-19 from Chest CT Images Using CNN with MLP Hybrid Model. *Stud Health Technol Inform*. 2021 Oct 27;285:288-291. doi: 10.3233/SHTI210617. PMID: 34734889.
- [11] K. F. Bushra, M. A. Ahamed, & M. Ahmad. Automated detection of COVID-19 from X-ray images using CNN and Android mobile. *Research on Biomedical Engineering*, 1–8. 2021. Advance online publication <https://doi.org/10.1007/s42600-021-00163-2>
- [12] A. Bernheim et al., "Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection," *Radiology*, vol. 295, no. 3, p. 200463, Feb. 2020, doi: 10.1148/radiol.2020200463.
- [13] K. el Asnaoui and Y. Chawki, "Using X-ray images and deep learning for automated detection of coronavirus disease," *Journal of biomolecular structure & dynamics*, vol. 39, no. 10, pp. 3615–3626, Jul. 2021, doi: 10.1080/07391102.2020.1767212.

- [14] S. Yadav, J. Kaur Sandhu, Y. Pathak, and S. Jadhav, “Chest X-ray scanning based detection of COVID-19 using deep convolutional neural network,” Aug. 2020.
- [15] S. Fang, “Review of Inception from V1 to V4,” Tech Hive. May 2020. <https://sheng-fang.github.io/2020-05-05-review-googlenet-v1-v4/>
- [16] L. Deng and D. Yu, “Deep Learning: Methods and Applications,” *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014, doi: 10.1561/20000000039.
- [17] V. Kurama, “A Review of Popular Deep Learning Architectures: ResNet, InceptionV3, and SqueezeNet,” *PaperspaceBlog*, 2019.
- [18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *CoRR*, vol. abs/1512.00567, 2015, [Online]. Available: <http://arxiv.org/abs/1512.00567>
- [19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.
- [20] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *CoRR*, vol. abs/1704.04861, 2017, [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [21] D. Berrar, “Cross-Validation,” in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 542–545. doi: <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>.
- [22] M. Hossin and M.N. Sulaiman, “Review on Evaluation Metrics for Data Classification Evaluation”, *International Journal of Data Mining and Knowledge Management Process*, Vol. 5. No. 2, April 2015