ICIC Express Letters
Volume **17**, Number **12**, December **2023**

ICIC International ©2023 ISSN 1881-803X
pp. 1303–1310

DEEP LEARNING WITH AUXILIARY LEARNING ATTENTION MECHANISM FOR OIL PALM FRUIT IMAGE RIPENESS CLASSIFICATION

Herman[1], Albert Susanto[1], Tjeng Wawan Cenggoro[2,3]
Dedy Ariansyah[3] and Bens Pardamean[1,3]

[1]Computer Science Department, BINUS Graduate Program
[2]Computer Science Department, School of Computer Science
[3]Bioinformatics and Data Science Research Center
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
{ herman005; albert.susanto001; dedy.ariansyah }@binus.ac.id; { wcenggoro; bpardamean }@binus.edu

Received December 2022; accepted March 2023

Abstract. *Currently the available oil palm fruit ripeness classification models are lacking in accuracy and reliability. As the consequence, the automation of oil palm fruit ripeness sorting is not pervasive in the industry. This is unfortunate, because oil palm is one of the leading commodities in agriculture industry, especially in Indonesia. To improve the accuracy, we propose a deep learning model with an addition-based attention mechanism as an oil palm fruit ripeness classification system. The result of this study shows that the proposed model improved the accuracy of the best previous deep learning model by 9.85%.*

Keywords: Oil palm fruit classification, Convolutional neural network, Visual attention, Computer vision

1. **Introduction.** As a tropical country, agriculture becomes one of the largest industries in Indonesia. Because of that, many researches have been conducted to improve the process of agriculture in Indonesia [1, 2]. Among the agricultural commodities, palm oil can be considered as the most important commodity in Indonesia. In fact, Indonesia is the largest oil palm exporter in the world [3]. Considering the importance of palm oil in Indonesia, an intensification of palm oil production can significantly boost the Indonesian national economy. For that purpose, improving the oil palm Fresh Fruit Bunch (FFB) harvesting workflow is one of the most promising approaches. Currently, most of the oil palm plantations use manual labor within the harvesting workflow. Therefore, developing an automatic harvesting system can significantly improve production productivity. One of the tasks that can be automated in the workflow is FFB sorting according to the ripeness. This task can be automated using a computer vision system that can classify FFB images based on their ripeness [4].

Noticeably, a robust automation of this task requires an advanced computer vision system, due to the complex appearance of a palm oil FFB. Several studies have been conducted to develop palm oil fruit classification system using handcrafted color features [5, 6, 7, 8]. However, such approach is known to be error-prone. A slight difference in illumination between the real environment and the training dataset can degrade the performance of this approach significantly. Thus, this approach is not ready to be deployed in the industry. Fortunately, this problem can be solved by using a computer vision system based on deep learning. This approach was started to be adopted for FFB ripeness classification, by employing deep learning architectures for general image classification with

DOI: 10.24507/icicel.17.12.1303

no modification [9, 10]. It is possible that modifying the architecture to be more suitable for oil palm FFB classification would result in a better performance.

Among the possible modification for deep learning architectures, attention mechanism is the most promising to be applied. It is defined as a module that lets the deep learning model strongly attend only part of the image to classify images, as illustrated in Figure 1. This module is potentially suitable for oil palm FFB classification, because the ripeness of the FFB can be determined only by looking at the color of the fruits, not the image in general. Thus, the contribution of this study is to design an attention mechanism module that can improve the accuracy of a standard deep-learning-based image classification model for oil palm FFB classification. The study is organized as follows. The background and related works are presented in Sections 1 and 2, followed by methodology in Section 3. Section 4 presents the results and discussion, and Section 5 concludes the findings of the study.

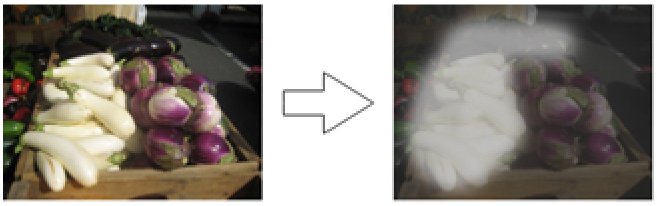

Figure 1. Eggplant detection attention

2. **Related Works.** Multiple researches have been conducted to develop deep learning models with attention mechanism. The first attempt to apply the attention mechanism into deep learning was done by Mnih et al. [11] for image classification. The attention was generated via recurrent neural network, which is applied afterward to generate caption from an image [12]. For image captioning, two types of attention were developed: a soft attention mechanism that generates continuous value as the attention and a hard attention that generates binary value as the attention. Inspired by this study, various attention mechanism modules were developed for computer vision cases such as image segmentation [13], place recognition [14], medical image analysis [15], and also general image classification.

Specifically for general image classification, the current most notable attention module is Squeeze and Excitation (SE) module. It injects a convolutional layer a capability to attend to its output feature maps channel-wise. In other words, it can be called as a channel attention. It successfully improved the accuracy of various standard Convolution Neural Network (CNN) for image classification on ImageNet dataset. SE is extended by Woo et al. [16] with the addition of spatial attention module on top of the channel attention. The extended module is called Convolutional Block Attention Module (CBAM).

Not only for computer vision, attention mechanism is also emerging for speech recognition [17]. It is even considered as a monumental breakthrough in natural language processing, with the formulation of self-attention [18], which gave birth to a new model called Transformer. It is currently the base of almost all popular models in natural language processing, such as Bidirectional Encoder Representations from Transformers (BERT) [19] and Extra-Long Neural Network (XL-Net) [20].

3. **Proposed Model.** This section discusses the proposed model named AuxNet, which uses DenseNet as the backbone network.

3.1. **DenseNet.** DenseNet [21] is a network that utilizes dense connections, whose idea is to concatenate the feature map of all previous layers to the subsequent layer in a block. The dense connections allow DenseNet to reuse previous feature maps, which ease the training process. DenseNet is also more parameter-efficient than ResNet [22], which connects only the output of the immediate previous layer. Figure 2 illustrates the inner working of DenseNet.

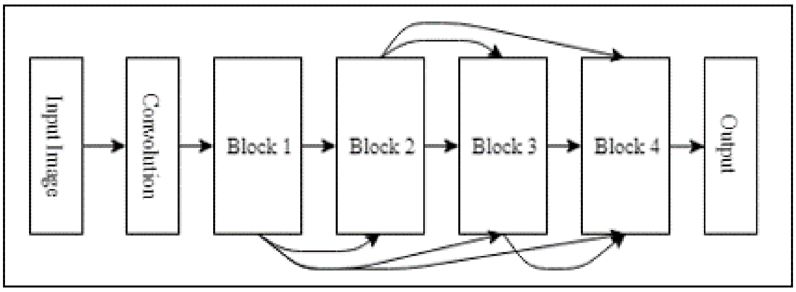

FIGURE 2. DenseNet architecture

3.2. **Auxiliary learning attention mechanism.** Attention mechanism is usually implemented as a learned feature map that is multiplied element-wise to the final feature maps of a CNN. To model attention with the learned feature map, it usually goes through a softmax or sigmoid activation function. Attaching an attention module only to the final feature maps limits the expressiveness of the model. However, attaching it to intermediate feature maps leads to an architecture that requires gradient from earlier layers to pass through softmax/sigmoid function multiple times. Because the magnitude of the gradients is decreased at every pass, the model is prone to heavily suffer from vanishing gradient problem.

To cope with this issue, we proposed an attention mechanism module that throws its output to an auxiliary classifier instead of passing it to the next layer. We call this attention mechanism as Auxiliary Learning Attention Mechanism (ALAM). Figure 3 illustrates

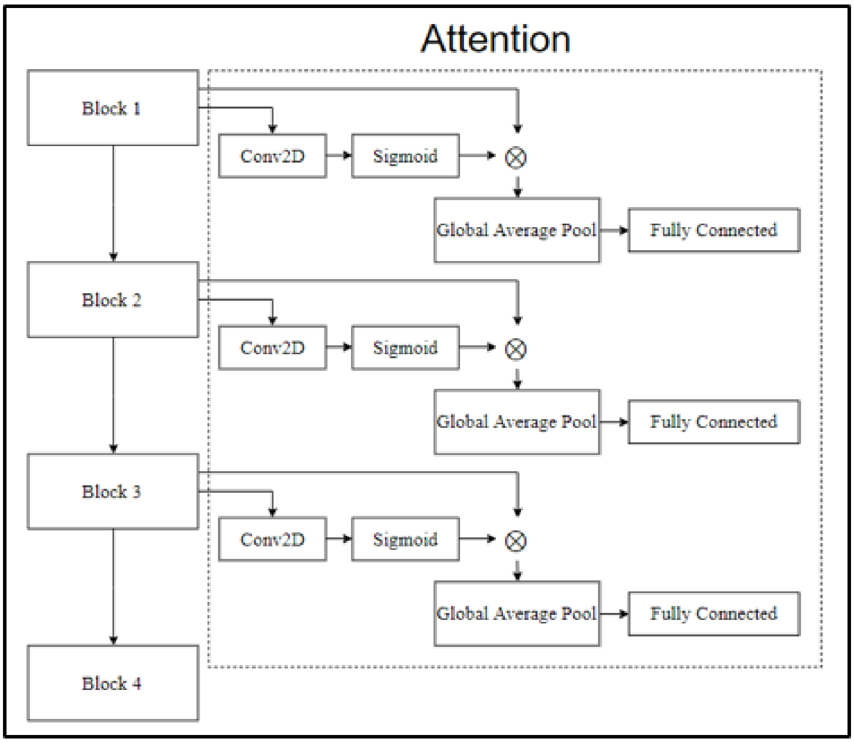

FIGURE 3. AuxNet

our proposed model, where the attention modules are attached to the output of each block in a DenseNet architecture. ALAM uses a single Conv2D that outputs the same tensor size as the input before going through the Sigmoid activation function which produces the attention area. Each tensor will retain its own attention as each channel focuses on each of their features. The attention tensor will then be multiplied element-wise with the original input before going through a Global Average Pool. The dot product will then be converted into a fully connected layer that will classify the image. The only block that did not use attention is the last one as the network does not continue to another block afterward. The proposed network that uses ALAM is referred as AuxNet (Auxiliary Network) for the rest of this paper.

To benchmark our proposed model, we compared it to a standard AlexNet that was used by Ibrahim et al. [9], a standard DenseNet, and a DenseNet with SE modules. All models used a pretrained ImageNet model to have benefits from transfer learning. The latter was used in the benchmarking experiment to test whether our proposed attention module is better than the state-of-the-art attention module. We employed the DenseNet with SE modules introduced by Yan et al. [23], which enables the use of a pretrained DenseNet.

In the benchmarking experiment, we used a dataset of 400 images oil palm FFB images, categorized into 7 ripeness classes. The distribution of images in the dataset is summarized in Table 1. The dataset was dividied into 3 groups for a typical CNN training procedure: training data (64%, 256 images), validation (20%, 80 images), and testing (16%, 64 images). To increase the variation in the dataset, all models were trained with Ten Crop, a data augmentation procedure that can increase a dataset variation ten-fold. It works by cropping the top-left, top-right, bottom-left, bottom-right, and center part of an image to produce five different images. These five images are also flipped horizontally to obtain more images.

Table 1. Classification of oil palm fruit

Name	Description	Total
BP	Ripening	16
BM	Raw	8
KM	Less Ripped	64
MKM	Almost Ripped	16
M	Ripped	96
MM	Perfectly Ripped	168
TM	Excessively Ripped	32
Total		400

4. **Results and Discussion.**

Table 2 shows the evaluation result of all models performance in the benchmarking experiment. It is clear that the models with DenseNet are superior from AlexNet. Interestingly, adding SE modules to the standard DenseNet reduced the performance. This phenomenon can be explained by the fact that an SE module passes its output to the next layer, which has been exposed to a softmax function. This design allows the gradients of earlier layer to pass through several softmax function, reducing the gradients magnitude for each pass. This leads to a vanishing gradient problem.

Contrarily to the SE module, adding ALAM module improved the performance of the standard DenseNet. This was possible by the fact that ALAM escapes its output to a classifier instead of passing it to the next layer. This guarantees any gradients within the model to only pass an ALAM module once.

Top-1 error rate and F1 Score are the evaluation methods used to measure the performance of all models proposed. All architecture was implemented with PyTorch under

Table 2. Test F1 Score

Model name	Accuracy	Precision	Recall	F1 Score
AlexNet	0.77	0.78	0.77	0.77
DenseNet	0.85	0.87	0.86	0.85
DenseNet + SE Layer	0.80	0.84	0.81	0.81
AuxNet	**0.87**	**0.88**	**0.87**	**0.87**

the same python environment on NVIDIA Tesla P100 and Tesla P4 GPU provided by NVIDIA – BINUS AI R&D Center. All models weight values were taken based on ImageNet transfer learning, so the model has already learned multiple features from general objects. The hyperparameters config can be seen in Table 3.

Table 3. Hyperparameter config

Hyperparameter	Value
Learning Rate (LR)	0.001
LR decay steps	Every 8 steps dropped for 10-1
Optimization	SGD
Batch size	8
Epochs	50
Pretrained	True
Freeze	False

The LR used is the most common and basic value that is widely used in every deep learning experiment. The learning steps have been tested both on every 8 and 30 epochs with no significant results changed; therefore, 8 epochs were used as the default for every model tested. The optimization used is Stochastic Gradient Descent (SGD), since it tends to work better on similar data between classes. The freeze was set to False so the network could continue learning new features on top of pretrained from ImageNet.

Using one of the current state-of-the-art architectures for the image classification task, DenseNet, with a deeper and complex model, it showed more promising results as seen in Figure 4. Therefore, all of the visual attention methods proposed were tested on DenseNet,

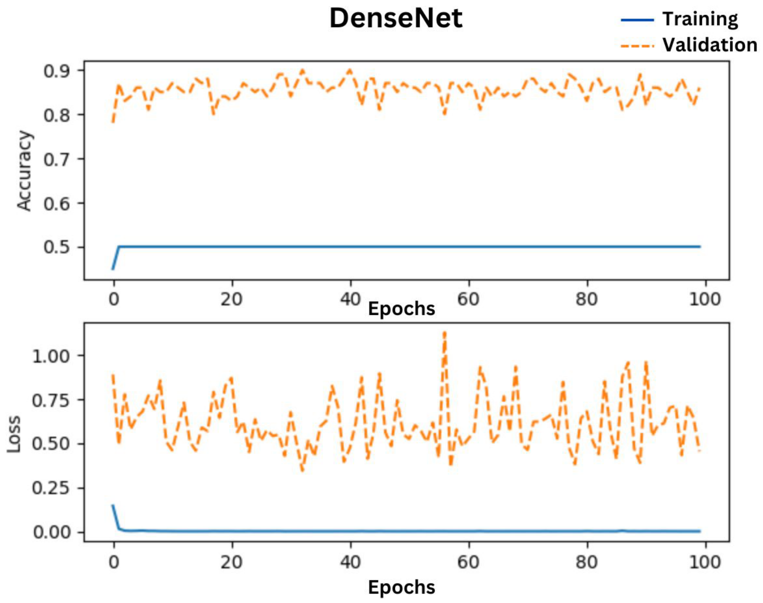

Figure 4. DenseNet training graph

particularly DenseNet121 during this research. The number of epochs used during training is 50 as most DenseNet and the proposed variation hit the lowest loss in the middle of training, in which the checkpoint was used for testing the result of each proposed model.

The method proposed in this research AuxNet uses the result of each block after going through a 3 × 3 convolution layer with Sigmoid activation function producing the same tensor size and channel depth before element-wise multiplication with the original block. The training result of AuxNet can be seen in Figure 5. The result in Table 4 is the lowest loss checkpoint from validation achieved from each model during the training. Since AuxNet consists of 4 fully connected layers that are used during backpropagation, it is natural to have a higher validation loss.

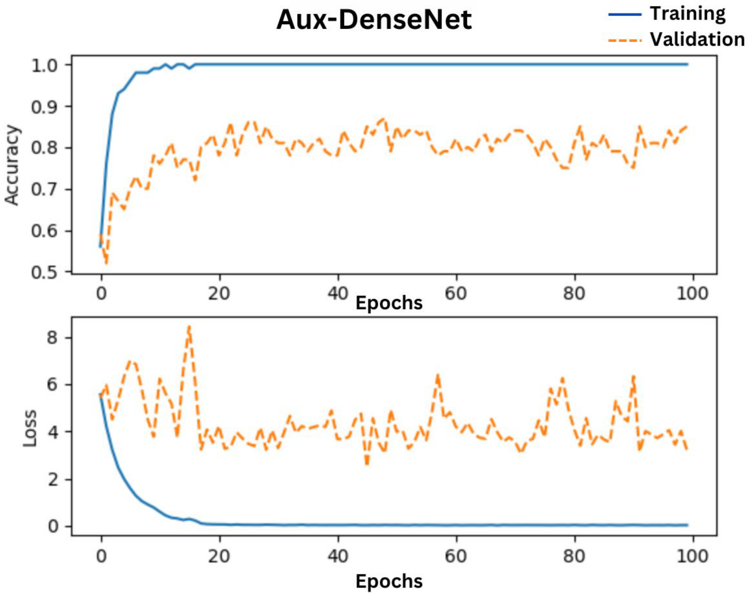

Figure 5. AuxNet training graph

Table 4. Training results

Model name	Lowest validation loss
AlexNet	0.9081
DenseNet	0.4467
DenseNet + SE layer	0.8904
AuxNet	2.4875

After training all of the models, all of them ran on testing data which consist of 64 images across all the class. The F1 Score of each model is shown in Table 2, AuxNet has proved to be able to increase the performance of DenseNet (DenseNet121). As this paper main contribution is AuxNet a confusion matrix will be used to show a more detailed performance between the model proposed and the highest second accuracy during the test model. Using AuxNet extension increases the base DenseNet precision, recall, F1 Score, and accuracy performance as seen in Table 2. More details of the performance of DenseNet121 and AuxNet can be seen in Tables 5 and 6, respectively.

Referencing from [24, 25], a confusion matrix composed of the binary result from DenseNet and AuxNet to the ground truth label was conducted. Each prediction was converted into 1 if it predicted correctly and 0 for the wrong classification. Then the binary result of both models was compared, if both were 1 then it is True Positive, else vice versa for True Negative. When both results were different while DenseNet predicted 1, it was considered False Positive, else False Negative if it was 0.

Table 5. DenseNet confusion matrix

		Prediction						
		BP	BM	KM	MKM	M	MM	TM
Ground truth	BP	36	0	0	0	0	0	0
	BM	0	12	0	0	0	0	0
	KM	0	0	86	20	0	31	7
	MKM	1	0	0	35	0	0	0
	M	2	0	0	0	216	8	2
	MM	7	0	8	0	18	358	5
	TM	1	0	0	0	10	11	50

Table 6. AuxNet confusion matrix

		Prediction						
		BP	BM	KM	MKM	M	MM	TM
Ground truth	BP	24	0	0	0	0	12	0
	BM	0	12	0	0	0	0	0
	KM	0	0	127	12	0	5	0
	MKM	0	0	12	24	0	0	0
	M	0	0	15	0	192	9	12
	MM	0	0	11	0	18	367	0
	TM	0	0	0	0	13	0	59

5. **Conclusion.** The attention generated from auxiliary learning could help the model learn better by outputting what it is being processed in the middle of the network during backpropagation. With the help of the attention block, the model could also produce a more focused dot product in which each feature has its important region rather than the overall features. The extension is able to increase the base model (DenseNet) accuracy, precision, recall, and F1 Score. The improvement over the DenseNet performance is 2%, 1%, 1%, and 2% respectively on accuracy, precision, recall, and F1 score. In the future, it is suggested to integrate ALAM on self-attention mechanism in a Visual-Transformers-based model.

Acknowledgment. The experiments in this study were executed using NVIDIA Tesla P4 and P100 supported by NVIDIA – BINUS AI R&D Center.

REFERENCES

[1] T. W. Cenggoro, A. Budiarto, R. Rahutomo and B. Pardamean, Information system design for deep learning based plant counting automation, *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*, pp.329-332, 2018.

[2] H. Sastrohartono, A. P. Suryotomo, S. Saifullah, T. Suparyanto, A. S. Perbangsa and B. Pardamean, Drone application model for image acquisition of plantation areas and oil palm trees counting, *2022 International Conference on Information Management and Technology (ICIMTech)*, pp.167-171, 2022.

[3] R. A. Pratama and T. Widodo, The impact of nontariff trade policy of European Union crude palm oil import on Indonesia, Malaysia, and the rest of the world economy: An analysis in GTAP framework, *Jurnal Ekonomi Indonesia*, vol.9, no.1, pp.39-52, 2020.

[4] H. Herman, T. W. Cenggoro, A. Susanto and B. Pardamean, Deep learning for oil palm fruit ripeness classification with DenseNet, *2021 International Conference on Information Management and Technology (ICIMTech)*, vol.1, pp.116-119, 2021.

[5] M. S. M. Alfatni, A. R. M. Shariff, H. Z. M. Shafri, O. B. Saaed, O. M. Eshanta et al., Oil palm fruit bunch grading system using red, green and blue digital number, *Journal of Applied Sciences*, vol.8, no.8, pp.1444-1452, 2008.

[6] N. Jamil, A. Mohamed and S. Abdullah, Automated grading of palm oil fresh fruit bunches (FFB) using neuro-fuzzy technique, *2009 International Conference of Soft Computing and Pattern Recognition*, pp.245-249, 2009.

[7] W. Ishak, R. Hudzari et al., Image based modeling for oil palm fruit maturity prediction, *Journal of Food, Agriculture & Environment*, vol.8, no.2, pp.469-476, 2010.

[8] N. Fadilah and J. Mohamad-Saleh, Color feature extraction of oil palm fresh fruit bunch image for ripeness classification, *The 13th Int. Conf. Appl. Comput. Appl. Comput. Sci.*, pp.51-55, 2014.

[9] Z. Ibrahim, N. Sabri and D. Isa, Palm oil fresh fruit bunch ripeness grading recognition using convolutional neural network, *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol.10, no.3-2, pp.109-113, 2018.

[10] Harsawardana, R. Rahutomo, B. Mahesworo, T. W. Cenggoro, A. Budiarto, T. Suparyanto, D. B. S. Atmaja, B. Samoedro and B. Pardamean, AI-based ripeness grading for oil palm fresh fruit bunch in smart crane grabber, *IOP Conference Series: Earth and Environmental Science*, vol.426, 012147, 2020.

[11] V. Mnih, N. Heess, A. Graves et al., Recurrent models of visual attention, *Advances in Neural Information Processing Systems*, pp.2204-2212, 2014.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *International Conference on Machine Learning*, pp.2048-2057, 2015.

[13] L.-C. Chen, Y. Yang, J. Wang, W. Xu and A. L. Yuille, Attention to scale: Scale-aware semantic image segmentation, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3640-3649, 2016.

[14] Z. Chen, L. Liu, I. Sa, Z. Ge and M. Chli, Learning context flexible attention model for long-term visual place recognition, *IEEE Robotics and Automation Letters*, vol.3, no.4, pp.4015-4022, 2018.

[15] Q. Tao, Z. Ge, J. Cai, J. Yin and S. See, Improving deep lesion detection using 3D contextual and spatial attention, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp.185-193, 2019.

[16] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, CBAM: Convolutional block attention module, *Proc. of the European Conference on Computer Vision (ECCV)*, pp.3-19, 2018.

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, Attention-based models for speech recognition, *Advances in Neural Information Processing Systems*, pp.577-585, 2015.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, pp.5998-6008, 2017.

[19] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp.4171-4186, 2019.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems*, pp.5753-5763, 2019.

[21] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Densely connected convolutional networks, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4700-4708, 2017.

[22] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[23] C. Yan, J. Yao, R. Li, Z. Xu and J. Huang, Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays, *Proc. of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp.103-110, 2018.

[24] B. Pardamean, *Dasar Bioinformatika Dengan R (Fundamentals of Bioinformatics with R)*, Graha Ilmu, 2017 (in Indonesian).

[25] B. Pardamean, A. Budiarto and R. Caraka, *Bioinformatika Dengan R Tingkat Lanjut (Advanced Bioinformatics with R)*, Graha Ilmu, 2018 (in Indonesian).