

# MULTI-FOOD RECOGNITION IN SINGLE TRAY BOX IMAGE WITH SCARCITY DATA USING CONVOLUTIONAL NEURAL NETWORK

Yuita Arum Sari<sup>1</sup>, Sigit Adinugroho<sup>1</sup>, Jaya Mahar Maligan<sup>2</sup>,

Muh. Arif Rahman<sup>1</sup>, Yusuf Gladiensyah Bihanda<sup>1</sup>

<sup>1</sup>Faculty of Computer Science, University of Brawijaya, Indonesia, yuita@ub.ac.id

<sup>2</sup>Agricultural Product Technology Department, University of Brawijaya, Indonesia

## ABSTRACT

Food recognition research is growing rapidly since it is able to be applied in some applications, including leftover estimation. In previous research we built prototype, named Smart Nutrition Box for measuring the leftover automatically. However, before calculating the remaining weight of food, we need to recognize the type of food itself, since blank compartment (non-food detection) negatively affects estimation. In this paper, we propose a feature of multi-food recognition in single tray box images using Convolutional Neural Network (CNN). Automatic cropping by using rectangular contour is applied. However, for further recognition, we only have small amount of data from our own dataset as training and validation images, so a technique to deal with scarcity data using image augmentation is used. Different type of activation function in the dense layers are tested: Rectified Linear Units (ReLU), Sigmoid, and Hyperbolic Tangent (Tanh). The result shows that ReLU activation function is the best with no dropout and reaches 90% of accuracy. If dropout is applied in dense layer, the Sigmoid activation function is better with slightly different of accuracy 89% in recognition. It concludes that CNN can be applied to recognize multiple food in single image tray box.

**Key words:** Food recognition, Scarcity image dataset, Convolutional Neural Network.

## 1. INTRODUCTION

Leftover food estimation and food recognition is becoming popular research nowadays, since every person needs to be aware with everything they consumed. This application also useful for people who manage their diet and control their appetite (Akhi *et al.*, 2018; Jiang *et al.*, 2020).

In the previous research, we also created a prototype called as Smart Nutrition Box to estimate leftover food (Sari *et al.*, 2020). However, there is a problem when we estimate food in the blank compartment of tray box image. It makes wrong estimation since the method is based on segmentation phase, thus, to minimize the error the following research is directed to determine food or non-food in single tray box image. If the compartment is detected as non-food it will be estimated as zero gram of food and if the compartment contains food, then continue to further process of estimation using leftover estimation algorithm (Sari *et al.*, 2019). The recognizing of multiple food or non-food in single tray box image can achieve decent performance of estimation (Sari *et al.*, 2020). In order to increase one of features in our prototype, recognizing multi-food in single tray box image is needed, so that we do not need to annotate the testing data of tray box images containing multi-food one by one. This part of algorithm is necessary for further stage of leftover estimation more precisely.

Extracting features by using conventional methods, such as color and texture has complex problem when determining the best feature to classify food (Arivazhagan *et al.*, 2010). If a set of food images extracted by color only, then it finds difficult to distinguish one type of food to another, since the different of food may have same color. It is also happened when using texture feature only, one type of food may have various texture feature, so that it does not fit to find the characteristic of food. The distance metrics is also be a parameter that should be considered since it can affect the recognition process (Suttapakti and Bumpeng, 2019). Therefore, in order to extract the features automatically, the development of food recognition using deep neural network is used.

CNN represents the feature for large scale of dataset adaptively (Krizhevsky *et al.*, 2012). However, in this research we only have small dataset called as scarcity data and it overcomes by using image augmentation (Perez and Wang, 2017). Image augmentation has possibility to enlarge dataset such as: rotating, flipping, cropping, etc. Based on some reasons above, we propose new framework to recognize multi-food image in single tray box image by using CNN and image augmentation at once. In Section 2 the detail of material and research method are explained, including dataset and the preprocessing of dataset until the technique of recognition using CNN. Then, the result is described in the Section 3 and the analysis also discussed in the Section 3. Eventually, the paper is concluded in the Section 4.

## 2. MATERIAL AND METHODOLOGY

### 2.1 Dataset

The number of data in this paper is very limited to be used in the model of deep neural network, CNN. So that, we implement data augmentation to boost the size of training dataset by creating more variations of images in the dataset (Chen and Chao, 2019). In this paper, there are 36 tray box images as testing containing of 12 tray box images in black, 12 images in grey, and 12 in white background with two different lighting of white scale. In each tray box, it has 4 compartments containing food and non-food (blank compartment). In this paper, we focus on how to recognize food in the non-blank compartment. We use two different menus named as Menu 1 and Menu 2 which is depicted in Figure 1.



**Fig. 1. The type of dataset**

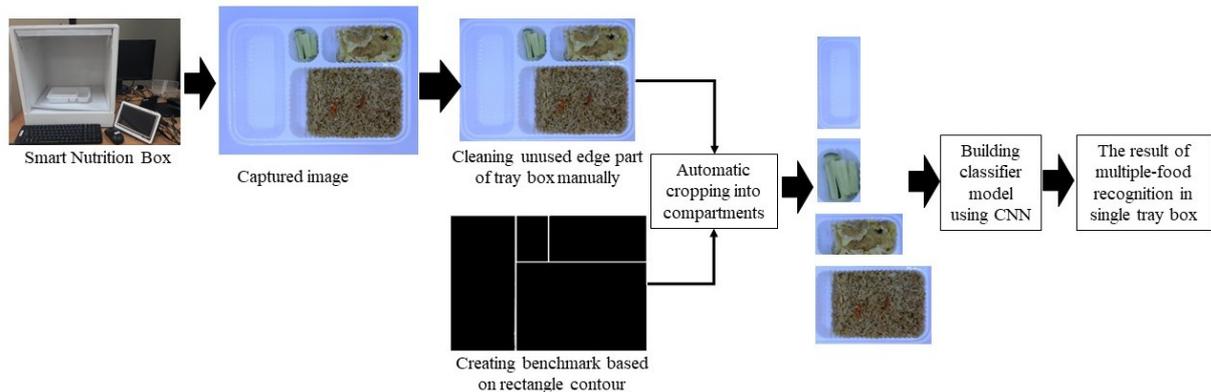
A set of train and validation images are used to build an ideal model to recognize an item of food in each compartment of tray box. In order to create our own dataset, we have 6 folders related to annotate the data. It consists of 5 food items and 1 label of non-food. The number of data in each label are imbalanced as referred in Table 1. There are 212 images belonging to 6 classes as train set and 53 images belonging to 6 classes in validation set.

**Table 1. The number of data in train set and validation.**

Label	Number of Data	
	training	validation
ayam goreng	10	2
mentimun	42	11
nasi goreng	49	8
nasi putih	10	2
telur	39	8
non-food	62	22

### 2.2 Research Methodology

Generally, the entire of methodology is presented in the Figure 2. The first phase is capturing tray box image using our previous prototype, Smart Nutrition Box. A tray box containing a menu which is consisting of several foods are placed inside prototype, exactly perpendicular to the bottom of web camera. All images are saved and cleaned manually in order to take only main part of tray box, so that we remove some unused edge surrounding a tray box. We create our own benchmark related to the pattern of our tray box, then apply automatic cropping algorithm to break down a single tray box image into compartments. The proposed automatic algorithm is utilized based on identifying rectangle contour in each part to obtain some information in terms of  $x$  and  $y$  coordinate, height and width. This information is used to acknowledge the size that should be cropped in other tray box images. To classify food items, we apply CNN to create model in both train and validation set images. Good model helps good prediction as well, so that choosing parameters among CNN algorithm is also important to be considered. The result is single tray box image with detail information of what item of food in its compartments.



**Fig. 2. Proposed method illustration**

### 2.2.1 Automatic Cropping Algorithm in Tray Box Image

In order to have equal dimensions among tray box images, thus we also propose the size calibration to apply cropping algorithm automatically (Sari *et al.*, 2020). By using cropping algorithm, a single tray box image is divided into four compartments automatically. An ideal benchmark as suitable for certain tray box is also considered. Algorithm 1 explains the detail process of automatic cropping in tray box images into compartments.

---

#### Algorithm 1. Automatic cropping algorithm from tray box image into compartments

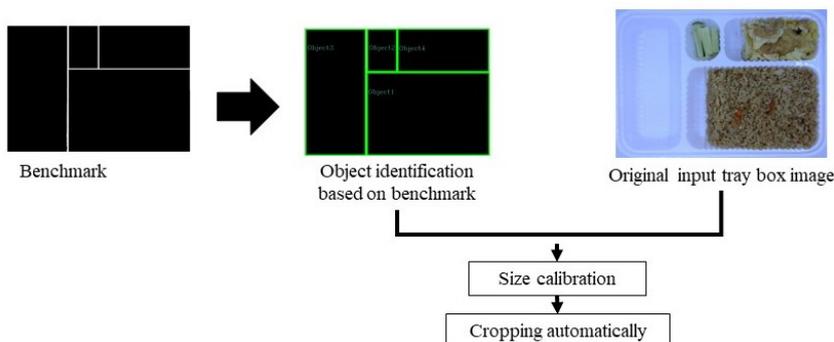
---

Input: Tray box image and benchmark

Output: Four cropped image called as compartments

Steps:

- 1) Coordinate and size identification from benchmark
    - a. Read benchmark image (black and white line)
    - b. Size identification from benchmark
  - 2) Finding contour information in order to get the detail size of each compartment
    - a. Color transformation from RGB image to grayscale image
    - b. Noise removal using Gaussian filtering with 15x15 kernel size
    - c. Get binary image using Otsu thresholding
    - d. Finding rectangular contour
    - e. Return list of  $x$  and  $y$  coordinates, height and width in each rectangular object
  - 3) Cropping tray box image into compartment
    - a. Read original image containing of food items in tray box
    - b. Size calibration of original images into benchmark's size
    - c. Get list of  $x$  and  $y$  coordinates from benchmark to be applied in original input of tray box image
  - 4) Cropped image as compartments from tray box image. Loop whole tray box images in a directory
- 



**Fig. 3. Automatic cropping algorithm into compartments**

### 2.2.2 Building Classifier Model using CNN with Image Augmentation

CNN is part of Neural Networks algorithm that is widely used as it is an effective method to recognize food in large scale of image dataset. In general, thousands of images are involved to build a pre-trained model. However, the problem is when capturing image from our own dataset, we have limited number of data, so that there is an

alternative to put image augmentation in order to be suitable for building a model since it can increase the number of datasets. Figure 4 shows the detail process of using CNN model with image augmentation. In this case, we use four types of image augmentation: shear range, zoom range, horizontal flip and fill mode. After that, constructing the model of CNN by using two phases of 2D convolution and max pooling is computed. In this paper, we only try to use 2 layers to extract the features. In the first convolution, we apply 16 kernels with 3x3 of its size and activation function ReLu. In the second convolution, we utilize two times of the kernels in the first convolution with the same kernel size. Both of max pooling have 2x2 kernel size in order to reducing image size by 50%. Based on the Figure 4, after 2 layers of extracting feature, then covert it into 1D of flattened matrix to be fed to the fully connected layers. Two dense layers is used, and we observe three types of dense activation functions to build different model of training. In the experiment of dense activation function, we apply three activation functions that are: ReLu, Sigmoid, and Tanh.

a) *ReLU activation function*

ReLU is the most used activation function in deep learning since it is computationally efficient to converge the network. The formula is given in the Equation 1.

$$R(z) = \max(0, z) \quad (1)$$

If  $z$  is less than 0 (means that is negative), so  $f(z)$  returns 0 as well as if  $z$  is equal to zero. However, if  $z$  is greater than zero, then  $f(z)$  is equal to  $z$ . ReLu is suitable for backpropagation when  $z$  is above zero, but when if the value of  $z$  is 0 or less than 0, then the backpropagation cannot well performed.

b) *Sigmoid activation function*

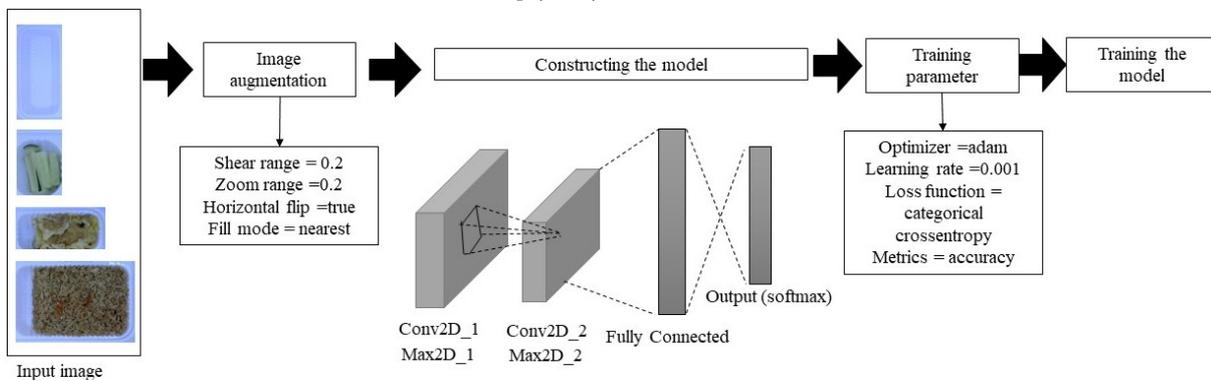
Sigmoid activation function is one of derivative functions which is necessary for training in neural network. The activation has S curve which has range between 0 and 1. It can reduce the problem in the gradient and normalizing output in each neuron. However, sigmoid achieve convergence slowly. Equation 2 describes the Sigmoid activation function.

$$f(x) = \frac{1}{1 + \exp(-x)} \quad (2)$$

c) *Tanh activation function*

It is similar with Sigmoid activation function, but Tanh is zero centered, so it is possible to have strong value of positive and negative as model input which has range between -1 and 1. Tanh activation function formula is illustrated in Equation 3.

$$f(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)} \quad (3)$$



**Fig. 4. CNN model for food classification**

The last part is tuning parameter in training model using Adam optimizer and it is proved that suitable for structured dataset (Wang, *et al.*, 2019). Learning rate is also applied to configure how quickly the CNN model learns a particular problem. In this paper we use 0.001 of learning rate. Besides, we apply loss function using categorical cross-entropy. Loss function is applied for giving estimation problem when training the model. Since, the dataset contains 6 classes of food detection and recognition, so multi-class cross entropy loss function is chosen. In order to provide the information related the how successful the training and validation model fits in the architecture, the accuracy metrics is adopted. After training parameters are set, then training the model using 10 steps per epoch, since we have about 200 train set of images and 20 batch size, so that 200 divided by 20 is 10. Next, we set 15 iterations in each epoch with 50 validation steps.

### 2.2.3 Food-Non Food Detection and Recognition using Tray Box Image as Testing Set

There are 36 testing data from tray box images with several conditions in different color of tray box. Train and validation set are used to get model from CNN architecture as depicted in the Figure 5. From testing set image, it is acknowledged per compartment and in the last step is concatenate whole compartment in one single tray box. Based on these techniques, it detects multiple food or non-food and recognize them as one menu in tray box.

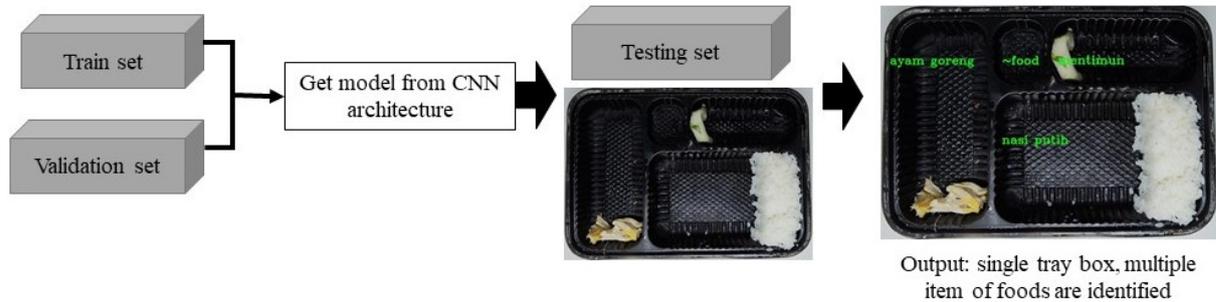


Fig. 5. Testing data tray box for each compartment

## 3. RESULT AND DISCUSSION

Three type of activation functions at dense layer are tested: ReLu, Sigmoid, and Tanh. In addition, to avoid some overfitting value, we also tested dropout and no dropout in the training process since we have imbalance number of data in each labels of training and validation. Dropout is also a technique of choosing randomly when selecting neurons during training process. The value of dropout is represented by using probability. We use small dropout probability: 0.5. The use of dropout in the experiments is depending on using dense activation function.

The accuracy of CNN model between training and validation with no dropout reach the best training using ReLu dense activation function with 98% accuracy of training and 96% of validation. The accuracy metrics between training and validation has slightly different, thus actually with no dropout the model is still good for recognition as stated in the Table 2. It is proven by accuracy of data testing achieves 90% of compartment images and 75% of tray box images (means that all compartment in a single tray box are correct to classify). Sigmoid is good as well but cannot overcome the performance of ReLu in this experiment which yields 85% of accuracy recognizing compartment images, although the accuracy of training and validation model are same in 98%. Sigmoid is stable dense activation function rather that ReLu and Tanh. If dropout do not include in dense layer, Tanh performance is sufficient good, but not the best comparing both of ReLu and Sigmoid. Figure 6 shows the example result of multiple food recognition in single tray box image using different dense of activation functions.

This condition is contrary when dropout is involved in the dense layer. In this case, using Tanh dense activation functions much better with raising accuracy of model in training and validation is 99% and 97%, respectively. The result of compartment image classification is the best among other dense activation functions which reaches 89% of accuracy. It is contrary when using ReLu dense activation function, the accuracy drops until 42% of tray box images consisting multiple food inside and 75% of compartment images. It means that using dropout does not have any impact to the recognition performance, except Sigmoid activation function which is converge in both situation: with dropout or without dropout. Figure 7 shows that using Tanh produces the best result to recognize multiple food in single tray box images.

Table 2. The accuracy model and data testing without dropout.

Dense Activation Function	Accuracy of Model		Accuracy of Data Testing	
	Training	Validation	Tray Box	compartments
Relu	98%	96%	75%	90%
Sigmoid	98%	98%	61%	85%
Tanh	91%	90%	39%	71%

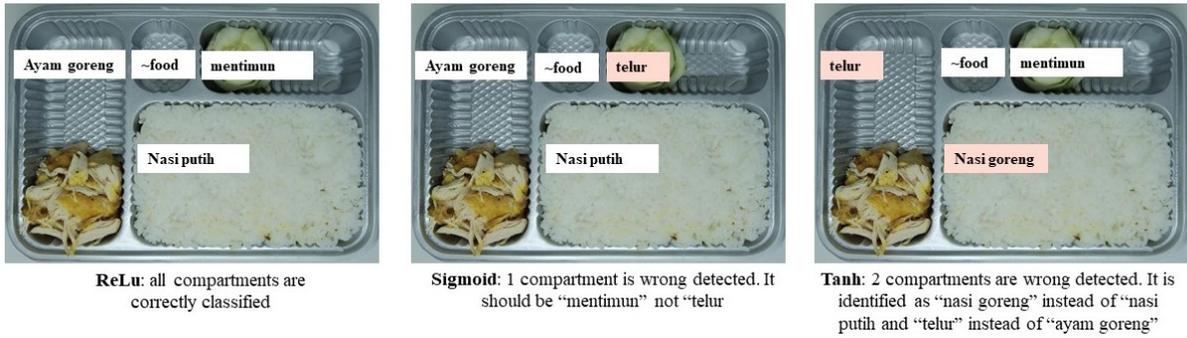


Fig. 6. The different result of ReLu, Sigmoid and Tanh without dropout testing

Table 3. The accuracy model and data testing with dropout.

Dense Activation Function	Accuracy of Model		Accuracy of Data Testing	
	Training	Validation	Tray Box	compartments
Relu	95%	96%	42%	75%
Sigmoid	97%	97%	50%	80%
Tanh	99%	97%	75%	89%

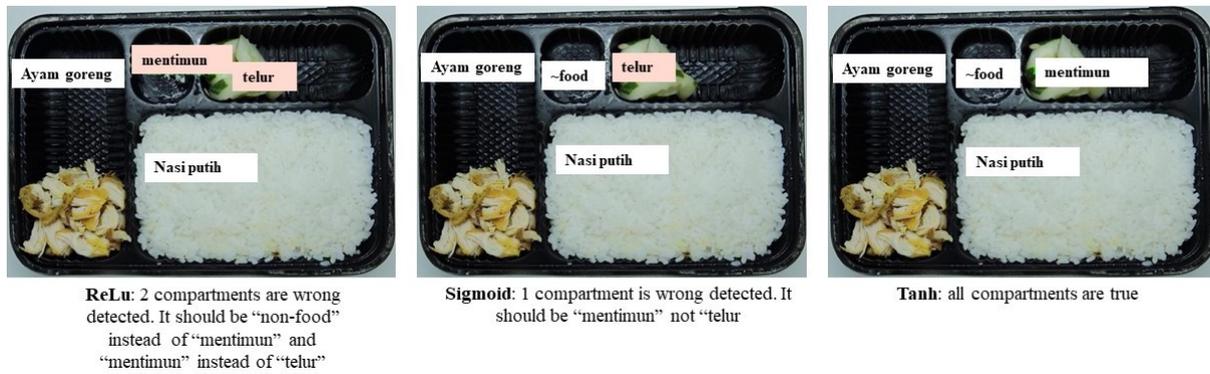


Fig. 7. The different result of ReLu, Sigmoid and Tanh with dropout testing

#### 4. CONCLUSION

Multi-food recognition in single tray box image with scarcity data using improved preprocessing and CNN is proposed. It has remarkable result with different experiment of dense activation function either using dropout and without dropout. From the entire experiment, without dropout is sufficient to classify each compartment image. The best result when using ReLu dense activation function which reaches 90% of accuracy in recognizing compartment images without dropout. When dropout is involved, then Tanh dense activation function is more suitable to use and, in this case, it reaches 89% of classifying compartment images in tray box.

The future works aims to improve the performance of preprocessing and classification since it needs to estimate the leftover in the tray box images after being consumed. Adding some training data or making it balance class is also treated, so that it expects to increase the accuracy of testing data from CNN model.

#### ACKNOWLEDGEMENT

The authors give thanks to Faculty of Computer Science (FILKOM), Brawijaya University, to support this research and also Ritsumeikan University which collaborates with FILKOM to compose this conference. This research is one of the consecutive results of research in development of prototype Smart Nutrition Box which is created between Computer Vision Research Group, Faculty of Computer Science and Agricultural Department of Technology, Brawijaya University.

## REFERENCES

- Akhi, A.B., Akter, F., Khatun, T., and Uddin, M.S. (2018). Recognition and Classification of Fast Food Images. *Global Journal of Computer Science and Technology*, 18(1), 7-13.
- Arivazhagan, S., Shebiah, R.N., Nidhyandhan, S.S., and Ganesan, L. (2010). Fruit recognition using color and texture features. *Journal of Emerging Trends in Computing and Information Sciences*, 1(2), 90-94.
- Chen, H., and Cao, P. (2019). Deep Learning Based Data Augmentation and Classification for Limited Medical Data Learning. In: *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Zhang, J.K., and Wang, Z.R. (Eds.), 300-303. IEEE: New York NY.
- Jiang, S., Min, W., Liu, L., and Luo, Z. (2020). Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition. *IEEE Transactions on Image Processing*, 29, 265-276.
- Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems 25 (NIPS 2012)*, Pereira, F., Burges, C.J.C., Bottou, L., and Weinberger, K.Q. (Eds.), 1097-1105. Curran Associates: Red Hook NY.
- Perez, L., and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.
- Sari, Y.A., Dewi, R.K., Maligan, J.M., Ananta, A.S., and dan Adinugroho, S. (2019). Automatic Food Leftover Estimation in Tray Box using Image Segmentation, In: *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, 212-216. IEEE: New York.
- Sari, Y.A., Dewi, R.K., Maligan, J.M., Maulana, L., and Adinugroho, S. (2020). Automatic Leftover Weight Prediction in Tray Box using Improved Color Lighting Component of Image Segmentation. *Journal of Southwest Jiaotong University*, 55(1), 1-18.
- Sari, Y.A., Maligan, J.M., Adinugroho, S., and Bihanda, Y.G. (2020). Multiple Food or Non-Food Detection in Single Tray Box Image using Fraction of Pixel Segmentation for Developing Smart Nutrition Box Prototype. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(3S), 132-136.
- Suttapakti, U., and Bunpeng, A. (2019). Potato Leaf Disease Classification Based on Distinct Color and Texture Feature Extraction, In: *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, Bao, V.N.Q. and Linh Trung, N. (Eds.), 82-85. IEEE: New York NY.
- Wang, Y., Liu, J., Mišić, J., Mišić, V. B., Lv, S., and Chang, X. (2019). Assessing Optimizer Impact on DNN Model Sensitivity to Adversarial Examples. *IEEE Access*, 7, 152766-152776.