Food Image Classification with Improved MobileNet Architecture and Data Augmentation

Sirawan Phiphiphatphaisit
PhD Student, Department of Information Technology
Faculty of Informatics, Mahasarakham University
Maha Sarakham, Thailand
61011261005@msu.ac.th

Olarik Surinta
Multi-agent Intelligent Simulation Laboratory (MISL)
Faculty of Informatics, Mahasarakham University
Maha Sarakham, Thailand
olarik.s@msu.ac.th

ABSTRACT
The real-world food image is a challenging problem for food image classification, because food images can be captured from different perspective and patterns. Also, many objects can appear in the image, not just foods. To recognize food images, in this paper, we propose a modified MobileNet architecture that is applies the global average pooling layers to avoid overfitting the food images, batch normalization, rectified linear unit, dropout layers, and the last layer is softmax. The state-of-the-art and the proposed MobileNet architectures are trained according to the fine-tuned model. The experimental results show that the proposed version of the MobileNet architecture achieves significantly higher accuracies than the original MobileNet architecture. The proposed MobileNet architecture significantly outperforms other architectures when the data augmentation techniques are combined.

CCS Concepts
• **Computing methodologies → Object recognition**
• **Computing methodologies → Neural networks.**

Keywords
Food Image classification; Convolutional Neural Network; MobileNet Architecture; Data Augmentation.

1. INTRODUCTION
Nowadays, people are becoming obese and overweight due to the imbalance between calorific intake and use. This increases the risk of other diseases such as diabetes, sleep apnea, acid reflux, and heart disease [12]. Nutritionists advise obese and overweight people to exercise and to monitor their daily consumption of calories [4]. Due to the assessment of calorie intakes into the body, Ege and Yani [3] proposed a multi-task convolutional neural network (CNN) method that allows the CNN architecture to learn from food calories, categories, ingredients, and cooking directions data. Furthermore, Myers et al. [13] presented a system that recognizes the contents of food from a single image, and then predict calories using the CNN based classifier. Then, people can estimate calories from food images.

In recent years, most research in food image classification has focused on hand-crafted features that consist of a color histogram [10,21], local binary pattern (LBP) [10,15], scale invariant feature transform (SIFT) [10], histogram of oriented gradients (HOG) [10,21], and speeded up robust feature (SURF) [2]. These hand-crafted methods are combined with machine learning algorithms to classify food images.

Due to the large-scale of food image datasets, researchers proposed to use deep learning algorithms to learn from the large-scale food image dataset such as the ETH Food-101 dataset which contains 101,000 images from 101 food categories; Food-256 dataset, a data set of 256 food categories with approximately 32,000 food images [2,5,7]. Yanai and Kawano [21] used a pre-trained model of AlexNet architecture for the feature extraction method. This method extracts 6,144 features from the image. In [5], the data augmentation techniques consist of brightness, contrast, saturation, and hue and are applied to food images before feeding to the Inception V3 network. Ming et al. [11] proposed the DietLens, which is a prototype of tracking dietary intake system for Singapore hawker food. The core architecture of the DietLens is the ResNet-50, which contains 50 convolutional layers and one fully connected layer and experiments on 87,470 images. The FoodNet [17], which is an ensemble deep neural network, is proposed to classify the Food-101 dataset. This network combined three well-known networks (AlexNet, GoogLeNet, and ResNet) as the ensemble network. The output of three networks and concatenate are passed to a fully connected layer to classify food images.

(A) (B)

Figure 1. Example of ETH Food-101 dataset. a) The apple pie category and b) the similarity shape between two categories of apple pie (first row) and Baklava (second row).

The challenge of food image classification is that food images from the same category are captured with different patterns, shapes, and perspective, accordingly to the people who take the image. For example, there are many objects such as forks and spoons, glasses, and bottles that appear in the image. For example of ETH Food-101 dataset, has many different apple pie images (that include other objects, patterns, shapes, and scenes) that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICISS 2020, March 19–22, 2020, Cambridge, United Kingdom
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7725-6/20/03…$15.00
https://doi.org/10.1145/3388176.3388179

appear in the apple pie category, as shown in Figure 1a). Even the similarity shape and pattern between the two categories of apple pie and Baklava, as shown in Figure 1b). These kinds of images can decrease the performance of the food image classification.

Related work: Hand-crafted feature extraction methods [14] are used in many image classification applications. In [15], two feature extraction methods consisting of a non-redundant local binary pattern (NRLBP) and the shape context descriptor of the interest points, called structure information are used to describe the local appearance information of food images. The achieved accuracy shows that the combination of the two features can improve classification performance. In [21], the first step uses, the color patches and RootHOG patches, (which is a square root of the L1 normalized HOG) to extract the data from the images. In the second step, the information from the first step is sent to a Fisher vector to encoding and used as the feature vector. This method achieved an accuracy of 65.3% on the UEC Food-100 dataset. In addition, Martinel et al. [10] presented the supervised extreme learning committee approach (ELM) to learning attributes of color, shape, texture, and local features. Then, the output of the ELMs is fed into the structured support vector machine (SVM) to classify food images. The performance achieved by this method is 55.89% and 84.34% on ETH Food-101 and UEC Food-100, respectively.

Nowadays, convolutional neural networks (CNNs), which are the most successful, and widely used for image classification problems [19]. Although, many CNN architectures can compute due to the large-scale images [19] and obtain very high accuracy [9,17]. In the area of food image classification, state-of-the-art CNN architectures such as AlexNet, GoogLeNet, and ResNet are proposed [17], although, the experimental results obtained with tem did not obtain high accuracy. Pandey et al. [17] invented a CNN-based ensemble network, called FoodNet architecture. This architecture consists of a fine-tuned model of AlexNet, GoogLeNet, and ResNet. The networks compute feature vectors and then concatenate all of the feature vectors, and a rectified linear unit (ReLU) used as a non-linear activation. Then, data is passed to a fully connected layer and the softmax function used to predict the output of the food image. The experiments showed that the FoodNet architecture obtained the Top-1 accuracy of 72.12% on ETH Food-101 and 73.50% on Indian food database. Also, the result was not good when the feature vector from the FoodNet architecture was fed into the SVM classifier.

As for the pre-trained model, In [21], the fine-tuning of the deep CNN pre-trained model based on AlexNet network, called DCNN was proposed to examine three food image datasets. The results showed that the fine-tuned DCNN achieved the Top-1 accuracy of 78.77%, 67.57%, and 70.40% on UEC Food-100, UEC Food-256, and ETH Food-101 datasets, respectively. The Inception networks [5,8] are proposed to address the food image classification. Lin et al. [8] presented the DeepFood network to recognize the food image for computer-aided dietary assessment. The DeepFood network, which is applied to an Inception module by adding 1x1 convolutional layers and then connected with two inception modules via an additional max-pooling layer. The best Top-1 accuracy results on UEC Food-256, UEC Food-100, and ETH Food-101 were 54.7%, 76.3%, and 77.4%, respectively. Hassannejad et al. [5] invented a deep network with 54 layers based on Inception V3 to classify three well-known food image datasets and achieved 88.28% on ETH Food-101, 81.45% on UEC Food-100, and 76.17% on UEC Food-256 datasets as top-1 accuracy.

Additionally, data augmentation is proposed to address the problem of insufficient data and to increase the performance of the image classification [1,22]. The data augmentation is also widely used in plant [18] and animal [16], and food [22] image recognition.

Contributions: In this paper, our main contribution is the use of the state-of-the-art deep convolutional neural network, called MobileNet architecture and our proposed MobileNet architecture is applied to recognize a challenging ETH food image dataset that contains 101 food categories.

In our proposed version, we reduce the number of parameters in the model by replacing the average pooling with the global average pooling (GAP) layers; then the overfitting is decreased. Subsequently, the batch normalization (BN), rectified linear unit (ReLU), and dropout layers, are utilized instead of the fully connected layers. Finally, the softmax layer is calculated. The results show that our proposed MobileNet architecture outperforms when compared to the original MobileNet architecture.

Moreover, we evaluate most effective data augmentation techniques to random creating images in the ETH food-101 dataset. We compared data augmentations and combined with the cropping image before passing to train the model. Also, the accuracy increased by approximately 5%. Finally, our proposed MobileNet architecture when combined with the data augmentation techniques outperforms the other methods.

***Paper outline*:** The paper is organized as follows: In Section 2, the MobileNet and the proposed MobileNet architectures are explained. In Section 3, the data augmentation techniques are presented. Experimental results are reported in Section 4. The last section is the conclusion and future work.

2. MOBILENET ARCHITECTURE

We used MobileNet architecture presented by Howard et al. [6] that is designed and based on depthwise separable convolutions to build a lightweight deep CNN that makes a model too small and reduces the computation time. The diagram in Figure 1a) illustrates the MobileNet architecture. Consequently, MobileNet can be implemented for several recognition problems such as object detection, face attributes, fine-grain classification, and landmark recognition.

2.1 Proposed MobileNet Architecture

Our proposed MobileNet architecture was as follows. First, we used the pre-trained model of MobileNet architecture. We decided to remove three layers, including the average pooling, fully connected, and softmax layers from the original network. Second, three extra layers; the global average pooling (GAP) layers, the batch normalization (BN), and softmax layers are attached. The main objective of our proposed MobileNet architecture is helping the network to train faster and achieving higher accuracy. Then, the dropout method is proposed to prevent overfitting. Also, the batch normalization layer helps the network to train faster. The activation function called the rectified linear unit (ReLU) is computed between the batch normalization layer and the dropout layer. After we applied the GAP layers instead of the average pooling, it shows that the parameters in the model are decreased, and impact directly on the size of the model. Finally, for training the proposed network, we used the fine-tuned MobileNet to train the network on the ETH Food-101 dataset. The proposed MobileNet architecture as shown in Figure 2b).

2.2 Depthwise Separable Convolutions Layer

The MobileNet architecture is computed based on depthwise separable convolutions (DS). The concept of decomposition of convolution called factorization is considered to factorize a standard convolution into a depthwise convolution.

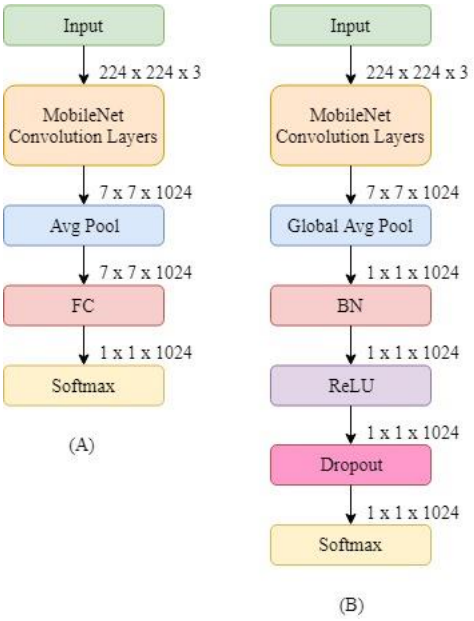

Figure 2. The architectures of the MobileNet. (A) the original MobileNet and, (B) the proposed MobileNet architectures.

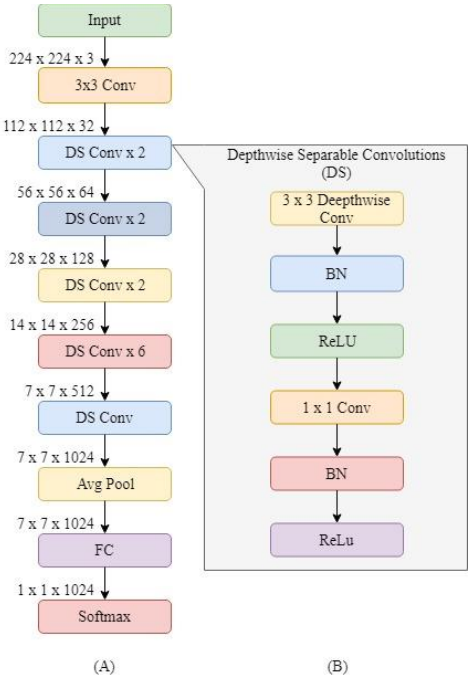

Figure 3. Illustration of the MobileNet architecture. (A) The overall MobileNet architecture and (B) an in-depth explanation of the DS layer.

After that, all depthwise convolution layers are computed with 1x1 convolution called a pointwise convolution, and then combined as the outputs to the next layer. The diagram in Figure 3a) shows the detail of the MobileNet that includes convolutional, depthwise separable convolutions (DS), average pooling, fully connected (FC), and softmax layers.

Figure 3b) shows an in-depth explanation of the DS layer consisting of depthwise convolution, batch normalization (BN), and rectified linear unit (ReLU), respectively.

3. DATA AUGMENTATION TECHNIQUES

Data augmentation is a technique to generate new training image data that relate to the same image. Many data augmentation techniques such as rotation, horizontal, vertical, flip, width shift, height shift techniques are applied to the image recognition problems and the accuracy performance is improved [22]. Samples of image augmentation are shown in figure 4. In this paper, the data augmentation techniques applied to our experiments consists of rescaling, rotation, width shift, height shift, horizontal flip, shear, and zoom.

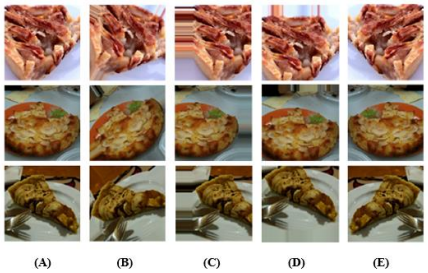

Figure 4. Example of the data augmentation images: (A) original, (B) rotation, (C) width shift, (D) height shift, and (E) horizontal flip images.

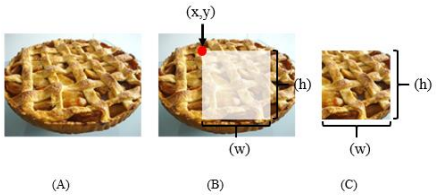

Figure 5. Illustration of the random cropping method. (A) Original food image, (B) random points (x,y) and crop sizes of the cropped image (w, h), and (C) the random cropping image used in training process.

Additionally, the image randomly changes to generate a new image in each training epoch, according to the range of the parameters.

Furthermore, random cropping is applied [20]. In this method, the position of points (x,y) are random, then it automatically crops and resizes to the target size, as shown in Figure 5. In this experiment, the size of the image is 224x224 pixel dimension.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 ETH Food-101 Dataset

In this paper, we evaluate the deep CNN architectures on the benchmark food image dataset. The real-world food images were collected by downloading from foodspotting.com website. The food images are a mix of eastern and western meals such as apple pie, Hamburger, Sashimi, Ramen, Peking duck. The challenging dataset consists of 101,000 food images from 101 food categories,

called the ETH food-101 dataset [2]. Examples of the food images are shown in Figure 6.

Figure 6. Sample real-world food images from the ETH Food-101 dataset.

4.2 Experimental Setup

Due to the large number of images in the dataset, we divided it into four subsets (Set I, Set II, Set III, and Set IV) sizes of 10,100 (randomly selected 100 images from each category), 20,200, 30,300, and 40,400 images to perform all of the experiments. Images in each subset were divided into training, validation, and testing sets of 70%, 10%, and 20%, respectively. For the training of the deep CNN architectures, we used the transfer learning with the following parameter settings: stochastic gradient descent (SGD) solver, batch size of 16, learning rate at 0.0001. We note that entire experiments were carried out using the TensorFlow platform running on Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz, 8GB RAM.

In the experiments, *firstly*, we used the original food images from the ETH Food-101 dataset to experimented with the MobileNet architectures in order to find the appropriate training epoch. *Secondly*, the first data augmentation called random cropping was employed. The program randomly cropped from a part of a food image and resize to the target size, which was 224x224 pixel dimension. *Thirdly*, the data augmentation techniques consisted of rescaling, rotation, width shift, height shift, horizontal flip, shear, and zoom applied according to the random parameters. Suddenly, the food images randomly change in each training epoch. *Finally*, the random cropping image and the data augmentation techniques are combined.

4.3 Experimental Results

We used 5-fold cross-validation in the training and testing phases. The accuracy and standard deviation are used to evaluate the performance of the deep CNN architectures on ETH food-101 dataset.

From the first experiment, it is essential to indicate that a huge number of food images can increase recognition performance. We set up the number of training to 50 epochs, which is similar to previous reports [1,17,23]. The accuracy of Set I with 10,100 images and Set IV with 40,400 images were significantly different. The accuracy results improved from around 42% to 57% when testted on the original MobileNet architecture. Moreover, the results improve from 46% to 67% when performed on the proposed MobileNet architecture, when accuracy increased by more than 10%, as shown in Figure 7. This clearly indicates that recognition performance is increased when using more food images.

We show the obtained results of second to fourth experiments using the proposed MobileNet architecture on four subsets of the ETH Food-101 dataset in Table 1. The table shows that the combination of the data augmentation and random cropping was the best approach in our experiments. This approach outperformed other methods with an increase of around 3-5% accuracy.

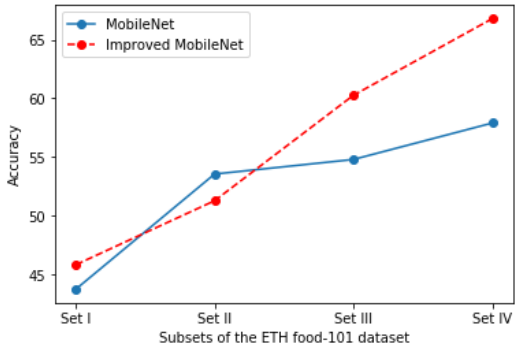

Figure 7. The performance of the MobileNet and proposed MobileNet architectures versus the different number of training samples (Set I - Set IV) on the ETH food-101 dataset.

Table 1. The performance results of food image recognition on four subsets on ETH Food-101 dataset using the approach MobileNet architecture

Methods	Subsets of the EHT Food-101 dataset			
	I	II	III	IV
Without data augmentation	45.84	51.29	60.26	66.78
Random cropping	45.79	55.82	59.52	67.44
With data augmentation	48.71	56.71	62.49	69.86
With data augmentation + random cropping	51.39	59.68	65.97	**72.59**

Table 2. Performances of the five different techniques on ETH Food-101 dataset

Method	The number of image per class	Accuracy (%)
Random Forest Discriminative Components [2]	1,000	50.76
Supervised Extreme Learning Committee [10]	1,000	55.89
Data Augmentation + MobileNet	400	57.90
Data Augmentation + Inception V3 [21]	1,000	70.41
FoodNet: Ensemble Net [17]	1,000	72.10
DeepFood [9]	1,000	**77.00**
Our proposed (Data Augmentation + MobileNet)	400	72.59

From the results in Table 2, the DeepFood architecture obtains the best performances on the ETH Food-101 dataset with an accuracy rate of 77%. Due to the computer used in the experiments, we decided to use the food image only 400 images per class to examine our proposed architecture. However, our proposed MobileNet architecture reached an accuracy of 72.59%. It is only 4.41% less than DeepFood architecture. As a result, our proposed MobileNet architecture outperforms the Random Forest

Discriminative Components [2], Supervised Extreme Learning Committee [10] and three deep CNN architectures; MobileNet, Inception V3 [21] and FoodNet [17].

In addition, the proposed MobileNet created a model size of 22.4MB, which less than the MobileNet architecture 10MB.

5. CONCLUSION

In this paper, we used the state-of-the-art MobileNet architecture on the food image dataset. We also described a MobileNet architecture, which was designed to address the overfitting problem. In this proposed MobileNet architecture, the number of parameters is decreased by applying the global average pooling (GAP) layers. Moreover, the batch normalization (BN), rectified linear unit (ReLU), and dropout layers are combined. Also, the last layer is the softmax. In addition, the data augmentation techniques are computed before transferring to the training process.

From the experimental results, to the best of our knowledge, we trained the MobileNet architecture according to the fine-tuned model. The proposed MobileNet architecture is competitive when compared to the original MobileNet architecture on the ETH food-101 dataset. We also demonstrated the impact of the data augmentation techniques; rotation, shift, flip, shear, zoom, and crop when implemented before assigning to the proposed MobileNet architecture to process. The best performance achieved when the combination of the various data augmentation techniques and the proposed MobileNet architecture.

In future work, we plan to construct the deep ensemble convolutional neural network (CNN) architectures, which are a combination of the state-of-the-art deep CNN architectures. We are interested in extracting the feature vector from the convolutional layers which may work better than individual deep CNN architecture.

6. REFERENCES

[1] Attokaren, D., Fernandes, I., Sriram, A., Murthy, Y. , and Koolagudi, S. 2017. Food classification from images using convolutional neural networks. *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2801–2806.

[2] Bossard, L., Guillaumin, M, and Gool, L. 2014. Food-101 -- Mining Discriminative Components with Random Forests. In *Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, Springer, Cham*. 8694, 446–461.

[3] Ege, T. and Yanai, K. 2017. Image-Based Food Calorie Estimation Using Knowledge on Food Categories, Ingredients and Cooking Directions. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017* (Thematic Workshops '17), 367–375.

[4] Fatehah, A., Poh, B., Shanita, S., and Wong, J. 2018. Feasibility of Reviewing Digital Food Images for Dietary Assessment among Nutrition Professionals. *Nutrients* 10, ,8 (July 2018), 1–12.

[5] Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I., Mordonini, M., and Cagnoni, S. 2016. Food Image Recognition Using Very Deep Convolutional Networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management - MADiMa '16*, 41–49.

[6] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv, abs/1704.04861*, 1–9.

[7] Kawano, Y. and Yanai, K. 2014. FoodCam-256: A Large-scale Real-time Mobile Food RecognitionSystem employing High-Dimensional Features and Compression of Classifier Weights. In *Proceedings of the 22nd ACM international conference on Multimedia (MM '14)*, 761–762.

[8] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., and Ma, Y. 2016. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *ICOST 2016. Lecture Notes in Computer Science, Springer, Cham*. 9677, 37–48.

[9] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y., Chen, S., and Hou, P. 2018. A New Deep Learning-Based Food Recognition System for Dietary Assessment on An Edge Computing Service Infrastructure. *IEEE Transactions on Services Computing*. 11, 249–261.

[10] Martinel, N., Piciarelli, C., and Micheloni, C. 2016. A supervised extreme learning committee for food recognition. *Computer Vision and Image Understanding.* 148, 67–86.

[11] Ming, Z., Chen, J., Cao, Y., Forde, C., Ngo, C., and Chua, T. 2018. Food Photo Recognition for Dietary Tracking: System and Experiment. In *MultiMedia Modeling*, 129–141.

[12] Must, A., Spadano, J., Coakley, E., Field, A., Colditz, G. and Dietz, W. 1999. The Disease Burden Associated With Overweight and Obesity. *JAMA* 282, 16 (October 1999), 1523–1529.

[13] Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., and Murphy, K. 2015. Im2Calories: Towards an Automated Mobile Vision Food Diary. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1233–1241.

[14] Nanni, L., Ghidoni, S., and Brahnam, S. 2017. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition.* 71, 158–172.

[15] Nguyen, D., Zong, Z., Ogunbona, P., Probst, Y., and Li, W. 2014. Food image classification using local appearance and global structural information. *Neurocomputing.* 140, 242–251.

[16] Okafor, E., Schomaker, L., and Wiering, M. 2018. An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals. *J. Information Telecommunication.* 2, 465–491.

[17] Pandey, P., Deepthi, A., Mandal, B., and Puhan, N. B. 2017. FoodNet: Recognizing Foods Using Ensemble of Deep Networks. *IEEE Signal Processing Letters.* 24, 1758–1762.

[18] Pawara, P., Okafor, E., Schomaker, L., and Wiering, M. 2017. Data Augmentation for Plant Classification. In *ACIVS, Springer, Cham*, 615–626

[19] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision.* 115, 211–252.

[20] Takahashi, R., Matsubara, T., and Uehara, K. 2018. Data Augmentation using Random Image Cropping and Patching for Deep CNNs. *ArXiv, abs/1811.09030*, 1–16.

[21] Yanai, K. and Kawano, Y. 2015. Food Image Recognition using Deep Convolutional Network with Pre-training and Fine-Tuning. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6.

[22] Yunus, R., Arif, O., Afzal, H., Amjad, M., Abbas, H., Bokhari, H., Haider, S., Zafar, N., and Nawaz, R. 2019. A Framework to Estimate the Nutritional Value of Food in Real Time Using Deep Learning Techniques. *IEEE Access.* 7, 2643–2652.

[23] Zheng, J., Zou, L., and Wang, Z. 2018. Mid-level deep Food Part mining for food image recognition. *IET Computer Vision.* 12, 298–304.