


RESEARCH

Open Access



Traditional food knowledge of Indonesia: a new high-quality food dataset and automatic recognition system

Ari Wibisono^{*} , Hanif Arief Wisesa, Zulia Putri Rahmadhani, Puteri Khatya Fahira, Petrus Mursanto and Wisnu Jatmiko

^{*}Correspondence:
ari.w@cs.ui.ac.id
Faculty of Computer Science,
Universitas Indonesia,
Indonesia, Kampus UI Depok,
Depok 16424, Indonesia

Abstract

Traditional food knowledge (TFK) is an essential aspect of human life. In terms of sociocultural aspects, TFK is necessary to protect ancestral culture. In terms of health, traditional foods contain better and more natural ingredients compared to the ingredients of processed foods. Considering this background, in this study, data acquisition and automatic food recognition were performed for traditional food in Indonesia. The food images were captured in a professional mini studio. The food image data were captured under the same light intensity, camera settings, and shooting distance from the camera. The parameters were precisely measured and configured with a light intensity meter, adjustable lighting, and a laser distance measurement device. The data of 1644 traditional food images were successfully obtained in the data acquisition process. These images corresponded to 34 types of traditional foods, and 30–50 images were obtained for each type of food. The size of the raw food image data was 53 GB. The data were divided into sets for training, testing, and validation. An automatic recognition system was developed to classify the traditional food of Indonesia. Training was performed using several types of convolutional neural network (CNN) models such as Densenet121, Resnet50, InceptionV3, and Nasnetmobile. The evaluation results indicated that when using a high quality dataset, the automatic recognition system could realize satisfactory area under the receiver operating characteristics (AUROC) and high accuracy, precision, and recall values of more than 0.95.

Keywords: Traditional food knowledge, High-quality dataset, Food recognition, Convolutional neural network

Introduction

Traditional food knowledge is the collective cultural wisdom of food systems that is passed down through generations [1]. Food systems involve the processes of production, processing, distribution, and consumption of food [2]. Hence, the sociocultural, economic, and health aspects of a specific ethnic group of a specific region can be inferred based on the traditional food knowledge. Traditional food knowledge helps in maintaining cultural diversity and ensuring food security for the community [1]. According to Sharif, the cultural identity of an ethnic community or region corresponds to the

traditional food [3]. In other words, traditional food represents the cultural diversity of a region, as it uniquely represents each ethnic community within the region. Therefore, it is essential to preserve the traditional food knowledge to maintain the diversity of the region. The food security in a region can be ensured by suitably managing the food systems. The Director of the Food and Agriculture Organization of the United Nations recommends implementing this concept based on the associated territory to rediversify the dietary intake of a community. Dietary diversity is specifically required for traditional diets [4]. Furthermore, localizing the food systems can support the local food suppliers in the agricultural sector and boost the economy of the region.

Indonesia has one of the most diverse cultural heritages in the world. The southeast Asian nation has more than 300 unique ethnic groups. As mentioned previously, the cultural identities of these ethnic groups are related to the traditional foods. Owing to the large number of unique ethnic groups in Indonesia, the number of distinct traditional food is also quite large. Despite the rapid urbanization of the country, the dietary preferences of both rural and urban residents still include traditional foods [5]. However, the dietary preferences of certain small ethnic groups that have moved to megacities such as Jakarta have shifted towards western food [5]. Recently, several efforts have been concentrated to ensure food security in Indonesia. Ensuring food security is essential to protect the diversity of traditional foods in Indonesia, as the appropriate food systems must be implemented for each region. However, according to a study conducted by the United Nations World Food Program (UN WFP), several regions in Indonesia are still categorized as “chronically food insecure.” One of the listed causes of this problem is that the food supply and availability threaten the food security in Indonesia [6]. This fact is also supported by an existing work [7]. In several regions of Indonesia, food is either under or oversupplied. The UN WFP recommends developing a surveillance system to realize food security and satisfy the nutrition requirements in Indonesia.

To the best of our knowledge, at present, no specific Indonesian food database is available that describes the traditional food knowledge of a specific region in Indonesia or can be used to realize a surveillance system for the food supply. Iqbal and Permadi mapped the nutritional values of Indonesian food. However, this study did not consider the food knowledge, in terms of how the food was produced and processed, among other factors [X1]. The Indonesian Ministry of Health attempted to achieve a similar goal using the website panganku.org, which primarily focuses on the composition of the nutrients inside the food. However, panganku.org does not specify certain necessary information about the food, such as its place of origin and grouping.

Considering these problems, the objective of this study was to create a new and scalable platform that can record the traditional food knowledge of Indonesian. The system was designed as a database to monitor the specific food systems for various Indonesian regions. The first phase of this system was to prepare a new Indonesian traditional food dataset to realize automatic food classification in Indonesia. This dataset was used to train the automatic classifier to classify Indonesian foods automatically. The dataset was prepared in a standardized manner in a professional laboratory setting. The standardization of data gathering was expected to enhance the accuracy of the classifier.

The dataset and automatic classification system can be used to prevent the loss of the traditional food knowledge. The dataset includes traditional foods from different ethnic

groups from several regions in Indonesia. Moreover, the system can be used as a surveillance system to monitor the supply of commodities in a given region, as recommended by the UN WFP. Thus, the system can help address the problem of threatened food security in Indonesia.

The remainder of this paper is organized as follows: The introduction section describes the general problem and the high level solutions proposed in this research. The subsequent section describes several related works on traditional food classification performed using various datasets and classification algorithms. The methodology section describes the process used to classify the traditional food data. The fourth section describes the automatic classification experiment setup and its results. The results and their analysis are discussed in this section. Finally, the conclusion summarizes the experimental findings of this paper.

Food datasets

Food datasets have been developed in several studies for various objectives. Most of these studies corresponded to the detection and recognition of the type of food to achieve the primary objectives, such as calorie counting and nutrition calculation. In several studies, food images were captured to record eating habits. Joutou and Yanai developed a 50-class food dataset from the web, consisting mostly of Japanese food with some western foods [8]. Subsequently, the authors extracted the features for each type of food and classified the data using MKL SVM. Wazumi et al. conducted a similar study in which Japanese food images were collected and classified to prepare a dietary log. However, instead of collecting the images from the web, the authors collected the food images from a university cafeteria, which were organized into 20 different classes. Moreover, each image in the dataset likely included various foods, as the images corresponded to the food placed on cafeteria trays [9]. Furthermore, Ciocca et al. performed multiple food classification. The authors acquired 3616 different food images from the university cafeteria, which were organized into 73 different classes. Subsequently, the authors classified the data by using a deep learning technique involving convolutional neural networks [10]. Zhu et al. collected 79 different classes of food, consisting of 20 to 50 different foods per class. Similar to in previous research, the image data contained various foods that were classified to assess the food and beverage intake. The food images were classified using the SVM classifier, with the color, texture, and SIFT features extracted from the dataset. In another study, researchers performed food recognition for dietary assessment [11]. American fast food images were acquired under restaurant and laboratory settings. In addition to the images, the authors also collected videos of the food. The dataset included seven classes of 101 different food images.

Furthermore, several researchers gathered food datasets to be implemented in systems to automatically determine the number of calories or food portions. In [12], 3000 different images were collected and classified into different types of foods. Subsequently, the system could automatically measure the calorie intake for the food. Similar efforts were made in studies [13] and [14]. The main focus of [13] was to classify food data from the ETH Food 101 dataset combined with the authors' Indian food dataset. The authors utilized ensemble deep networks, consisting of the AlexNet, GoogleNet, and ResNet frameworks. The ETH Food 101 dataset consists of 101 classes of food, and the authors'

Indian food dataset, extracted from the web, consists of 50,000 images classified into 50 classes.

Furthermore, in another study [14], a complete system to measure the calories of several foods in a single image was developed. The authors used the faster R-CNN to classify images from the web and the UEC Food 100 images, which primarily correspond to Japanese food. Anthimopoulos et al. classified food images to measure the carbohydrate (CHO) intake to develop a system for Type 1 diabetic patients. The image dataset consisted of 5000 different images acquired from the web. The authors utilized various classifiers, such as ANN, SVM, and random forest.

In various studies, images of local food were acquired to create a dataset. Phat et al. collected 2315 images of Vietnamese food divided into five different classes [15]. The images were obtained from the web. The authors used convolutional neural networks with handcrafted feature extractors, such as SIFT, SURF, and HOG. Similarly, in studies [16] and [17], Thai food was classified using convolutional neural networks and the modified visual geometry group (VGG) 19 network, respectively. In study [16], the authors developed a dataset via participants using a smartphone. The dataset included 3960 images divided into 11 different classes. In addition, in study [17], 7632 food images were collected from the web and divided into 11 different classes.

In this research, the data were acquired in a professional laboratory setting, similar to in study [11]. The literature review indicates that in nearly none of the existing studies were the image datasets acquired in a standardized environment. Therefore, in this work, the data acquisition process was standardized to ensure the same level of quality between each image in the dataset. The data for Indonesian traditional food were acquired. In recent literature [14, 18], deep learning approaches have been used to develop the classifier. Therefore, in this work as well, the data were classified using four different deep learning models: ResNet50, InceptionV3, Densenet121, and NasNetMob.

Proposed method

The methodology for this research is illustrated in Fig. 1. The investigation involved four main processes: acquisition of traditional food image data, preprocessing, development of automatic recognition model, and model evaluation. In the initial process, we captured the food images using professional photo studio equipment. High-quality food images were recorded professionally, and automatic classification of Indonesian food images was realized.

The images were obtained under identical settings for the light intensity, camera parameters (ISO, aperture, and shutter speed), object position, and camera position. In the preprocessing stage, the RAW files captured from the camera were edited to ensure that all the food images in the dataset had the same exposure and size. After ensuring that the data had identical exposure, training was performed using the food image dataset. A total of 1644 food images were obtained in the data collection process. The data consisted of 34 types of traditional foods, with 30–50 pictures for each type of food. The total size of the raw data for the food image dataset was 53 GB. The data.

were divided into training, testing, and validation data. Specifically, 70% of the data was used as training data, 20% was used as testing data, and 10% was used as validation data. As mentioned previously, the training on food image datasets was realized



using several kinds of deep learning network models such as Densenet121, Resnet50, Inception-V3, and Nasnetmobile. The evaluation results were evaluated considering the metrics of AUROC values, accuracy, precision, recall, F1 score, and model training time.

Acquisition of professional food images

As mentioned previously, the dataset consisted of 1644 images of traditional food. Several sample images in the processed food dataset are shown in Fig. 2. The traditional food considered in this research originated from various parts of Indonesia. Specifically, five, twenty, three, five, and three types of the food originated from the island of Sumatera, islands of Java-Bali-Lombok, Kalimantan, Sulawesi, and Papua, respectively. Thirty to fifty images were obtained for each food type. All the images in the dataset had consistent shooting angles, albeit with different augmentations. For every food image, the camera was set in two different angles, 45° and 90°. Every image in the dataset was obtained under identical shooting conditions (same exposure and ambience). The total size of the food image dataset was 53 GB.



The data were not collected using search engines. Instead, the food data were acquired by delivering each food item to the studio lab to ensure consistency in capturing the pictures. Each food item was treated equally during the acquisition processes. Therefore, the resulting images were obtained under identical conditions for the camera sensor. The

distance of the food to the camera was measured using a laser distance sensor placed 3 m from the front side and 1 m from the top side. The light intensity during shooting was 1250 lumens. Moreover, the camera settings were set to the same value to ensure that each type of food had the same exposure during imaging. The camera was used in the manual mode, with the aperture, ISO, and shutter speed set to $f/8.0$, 100, and $1/50$ s, respectively. This acquisition method is unconventional, as in general, almost all datasets are obtained with the camera in the automatic mode, which results in different exposures owing to the camera automatically changing the settings for each image capture.

Preprocessing

During data acquisition, the file generated by the camera is a RAW file and not a jpeg file. Standard camera settings produce JPEG files captured from camera sensors. However, the jpeg file does not fully represent the conditions captured by the sensor because an automatic process is conducted internally by the camera processor to ensure that the resulting image matches the shooting conditions. Moreover, the JPEG file is a lossy file, which means that certain captured results are ignored, and the information is not entered into the file to reduce the file size. Unlike jpeg files, raw files record all the pixel conditions of images captured by the camera sensors with a higher dynamic range than that of jpeg files. The resolution for each shot is approximately 4864×3868 pixels and 240 dpi.

At the preprocessing stage, each image was subjected to the same preprocessing treatment. If a jpeg file is used, the enhancing and preprocessing are done automatically by the camera. Jpeg images from the camera lead to a different exposure for each image. The size of each raw file ranges from 21–30 MB in contrast to the jpeg file, whose size is approximately 2–3 MB. The depth per inch (DPI) for the raw camera and smartphone camera images is approximately 240 DPI and 70 DPI, respectively.

Each captured image in the dataset (in raw form) was obtained under the same exposure. Therefore, all the preprocessed images had the same capture quality, ambient lighting, position and augmentation, thereby leading to a high quality image dataset. Subsequently, each image was cropped. The cropped image contained only the actual food. The cropped part was considered as the captured image of the food, and most images included a small cross-section, because certain parts of the food likely spilled over from the container.

Model development

To realize automatic recognition, the CNN model was used for automatic feature extraction. We tested several CNN network approaches for traditional food, namely, depth-based CNN principles and multipath based CNN.

The CNN architecture has been constantly improved since 1989, in terms of parameter optimization, structural changes, and regularization. However, the most significant contribution to the implementation of CNNs is the development of blocks and restructuring of the processing unit. The development trend of CNNs includes the development of more in depth or more diffused architectures. The improvement of CNNs can be categorized as spatial improvement, depth improvement, and multipath improvement.

In this work, depth-based CNN and multipath-based CNN were evaluated. Both the CNN methods have different operation mechanisms. The depth-based CNN is deeper and more efficient compared to shallow networks in preparing representations. However, these CNNs encounter the gradient vanishing problem. In contrast, in multipath-based CNNs, the path attempts to correct the vanishing gradient problem by creating a gradient that can access the lower layers. Our evaluation indicated that multipath-based CNN (Densenet-121) can achieve a higher AUROC compared to that of the depth-based CNN (Inception-V3).

Depth-based CNN

According to the CNN principle, a deeper network is better at target function estimation. This method increases the feature representation and nonlinear mapping. In theory, a deep network is more efficient compared to shallow networks in preparing representations. Csaji et al. considered a single hidden layer to be sufficient to perform several functions; however, this approach was not feasible owing to the large computations [19]. Bengio et al. suggested that deep networks may have the same representation capabilities, albeit with more efficient computing capacities [20, 21]. Inception and VGG, which represent depth based CNNs, exhibited a high performance in the 2014-ILVR computation [22–25].

ResNet ResNet was proposed by He et al. [26], and this network is a continuation of the deep network. ResNet can implement a CNN that is 152 layers deep. The architecture of the residual ResNet is 20 and eight times deeper than that of the original AlexNet [27] and VGG [28], respectively. He et al. demonstrated that the ResNet with 50,101,152 layers can achieve more accurate results compared to the results obtained using 34 layers. During testing, the Resnet demonstrated a 28% increase in the accuracy when using COCO 114 datasets. This finding proves that depth is an essential parameter to realize image recognition.

Inception-V3 Szegedy et al. proposed Inception-V3, which is an improved version of Inception V1/2 [25]. The main concept of Inception-V3 is to reduce the computation of a deep neural network without reducing the generalization ability by changing large filters (5×5 and 7×7) to small and asymmetrical filters (1×7 and 1×1 , respectively). This process narrows the filter before it is fed into a larger filter [25]. These filters make the convolutional processes similar to those of cross-channel correlation [29].

Inception-V3 utilizes a 1×1 convolution operation that maps the input data into 3 or 4 separate spaces smaller than the input space. The mapping correlations are performed in smaller 3D spaces with 3×3 or 5×5 convolutions. In the Inception-ResNet, Szegedy et al. combined the capabilities of residual learning and inception blocks [26, 29, 30]. At this stage, the residual connection replaces the concatenation filter. Szegedy et al. also demonstrated that training with a residual connection significantly accelerates the training of inception networks.

Multipath based CNNs

Deep network training is complex. Deep CNN can be used to realize the efficient computing of complex tasks. However, deep networks experience performance degradation in the form of gradient vanishing problems, which result in increased errors during training and testing [31–33]. However, it this phenomenon is not caused by overfitting [30, 34]. The concept of multipath or cross-layer connectivity can be used to address this problem [35–38]. The cross-layer network connectivity is divided into several blocks. The path attempts to correct the vanishing gradient problem by creating a gradient that can access the lower layers. To this end, several shortcut connections are developed, including zero padded, proconnection based, dropout, and 1×1 connections.

ResNet To address the problems in deep networks, He et al. proposed the use of ResNet. ResNet is a CNN architecture with a depth based CNN that bypasses the pathways [21]. This network was inspired by highway networks [35]. ResNet can be mathematically formulated as in Eqs. (1) and (2).

$$g(xi) = f(xi) + xi. \quad (1)$$

$$f(xi) = g(xi) - xi. \quad (2)$$

In Eq. (1), $f(xi)$ is transformed into a signal, where xi is the original input. The first input xi is added to $f(xi)$, thereby bypassing the pathways. In Eq. (2), $g(xi)$ performs residual learning operations. ResNet introduces shocut connections to exhibit cross-layer connectivity. Each gate in ResNet has independent data and parameters. In highway networks, when the shortcut to a gate is closed, the layers act as a residual function [35]. However, in ResNet, the residual information is always transmitted, and the shortcut identifier is never closed. The shortcut connections increase the convergence speed of a deep network and mitigate gradient diminishing problems.

Densenet Densenet is a continuation of the highway network and ResNet. Densenet can solve the vanishing gradient problem [26, 35, 36]. ResNet explicitly stores information through additive identity transformations, wherein many layers contain little to no representative information. Moreover, ResNet has a large weight for each layer. To solve this problem, DenseNet utilizes the cross-layer connectivity concept, albeit in a different manner. Specifically, Densenet connects the intermediate layers by using the feedforward method. In this case, the feature map from the previous layer is used as the input for all the subsequent layers.

Densenet connections are represented as $l(l+1)/2$, where l represents the connection between the layers in a traditional CNN. This aspect results in cross-layer depth wise convolution. Densenet does not add feature layers but concatenates the features of the previous layers. Therefore, the network can distinguish the information that must be stored and the information that must be added. Densenet has a simpler layer structure compared to that of the ResNet. However, Densenet has more parameters owing to the addition of the feature maps. The flow of information from each layer

through the gradient and loss function improves the process of delivering information on the network, leading to a positive regularization effect that reduces the overfitting during training.

Nasnet Nasnet helped in the realization of a search space known as the neural architecture space (NAS). The network child is trained to converge and enhance the accuracy in a held out test architecture. The results are used to update the controller so that the controller can have the optimal architecture. The building blocks of NASnet are repetitive blocks consisting of a combination of convolution filters. The convolution filters must be selected carefully to generate accurate results. NASnet can be used to develop an RNN controller to predict generic convolution cells, which can manage the inputs having several spatial dimensions and filter depths. NASnet has a top 1 accuracy of 74%, which is 3.1% better than that of state of the art models for mobile platforms [38].

Model evaluation

After classifying the images of the traditional food, the performance of the classification methods is evaluated considering the average area under receiver operating characteristic (AUROC), accuracy, precision, recall, and F1 score. We evaluated the AUROC results from every food class and obtained the average of those values. The accuracy was calculated using Eq. (3).

$$Accuracy = \frac{Correctly\ classified\ data}{Total\ data}. \quad (3)$$

Moreover, we obtained the precision, recall, and F1 score using Eqs. (4), (5), and (6), respectively.

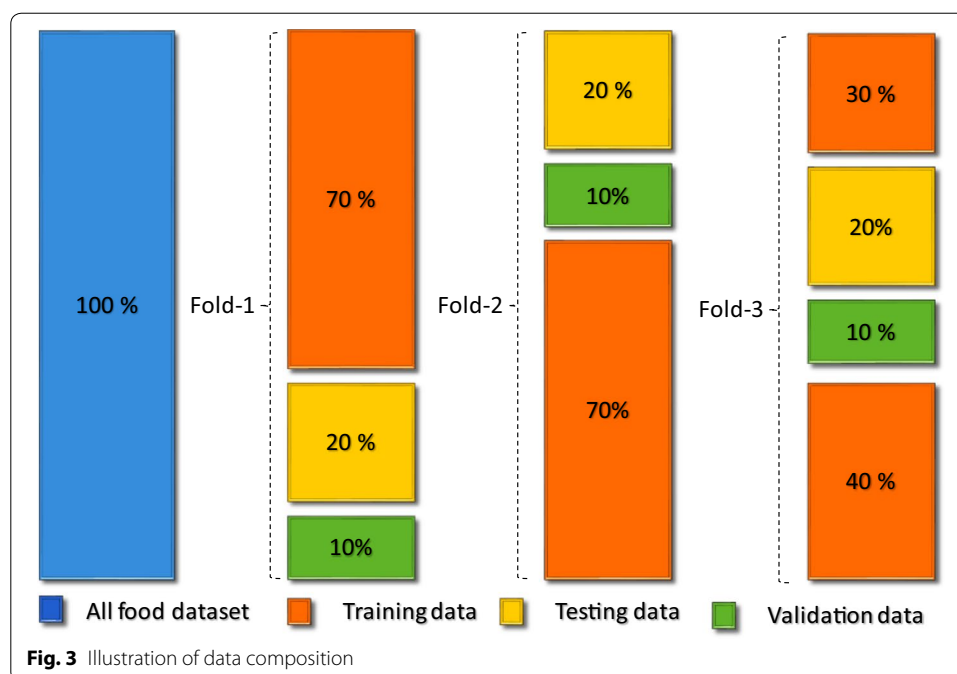
$$Recall = \frac{\sum_{i=1}^l tp_i}{\sum_{i=1}^l tp_i + fp_i}. \quad (4)$$

$$Precision = \frac{\sum_{i=1}^l tp_i}{\sum_{i=1}^l tp_i + fn_i}. \quad (5)$$

$$F1 = 2 \frac{Precision * Recall}{Precision + Recall}. \quad (6)$$

where tp_i is a true positive, fn_i is a false negative, and fp_i is a false positive. The F1 score was calculated using Eq. (6). In addition to the model accuracy, we also measured the execution time for each training model. The training model time was measured from when the image dataset was loaded up until 50 epochs in the training process. The workstations were used exclusively during training, and no other processes were conducted during training.

The data acquisition, preprocessing, and model development were performed consecutively. To evaluate the result, we assessed the traditional food image dataset obtained



using a few deep learning models based on a few metrics (AUROC, F1 score, and model training time).

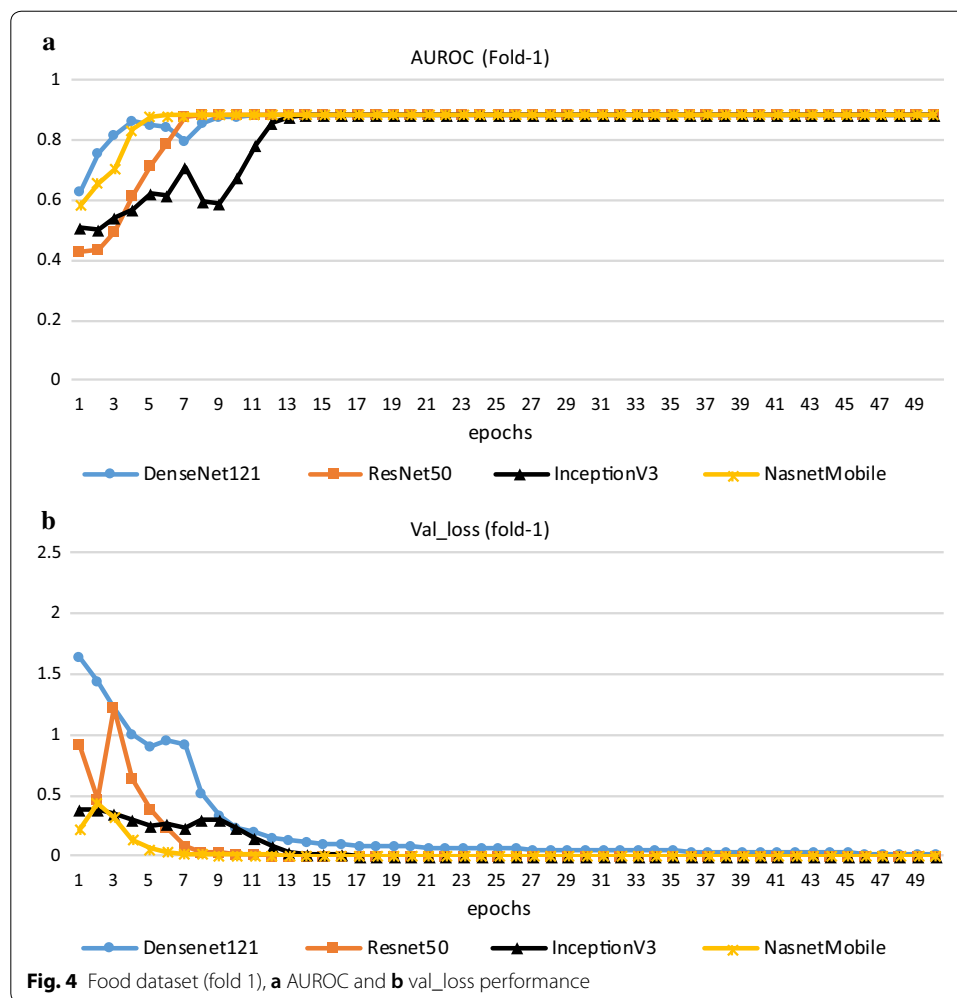
Results and discussion

This section describes the analysis of the obtained traditional food dataset. Several methods were used to divide the data into training data, testing data, and validation data. As described in the next subsection, we measured and analyzed the performance of several deep learning methods evaluated considering the dataset. The evaluation metrics included the AUROC value, F1 score, precision, and recall.

The limitation of this work corresponds to the lack of existing research on complete food classification in Indonesia. Several studies presented the result of small datasets for Indonesian food classification. However, the complete reference for the variations of food types representing several regions in Indonesia is not available. Moreover, in previous studies, the dataset was obtained from the internet or the food images only had a small sample size.

Data preparation and analysis

The data from the preprocessing stage was divided into three parts, namely, training data, testing data, and validation data. The training data was used by the model to realize supervised learning. The validation data was used to validate the supervised learning during training. The testing data was used to test the model after training. The composition of data was as follows: 70%, 20%, and 10% of the data were used as the training data, testing data, and validation data, respectively. The composition for each fold is illustrated in Fig. 3.

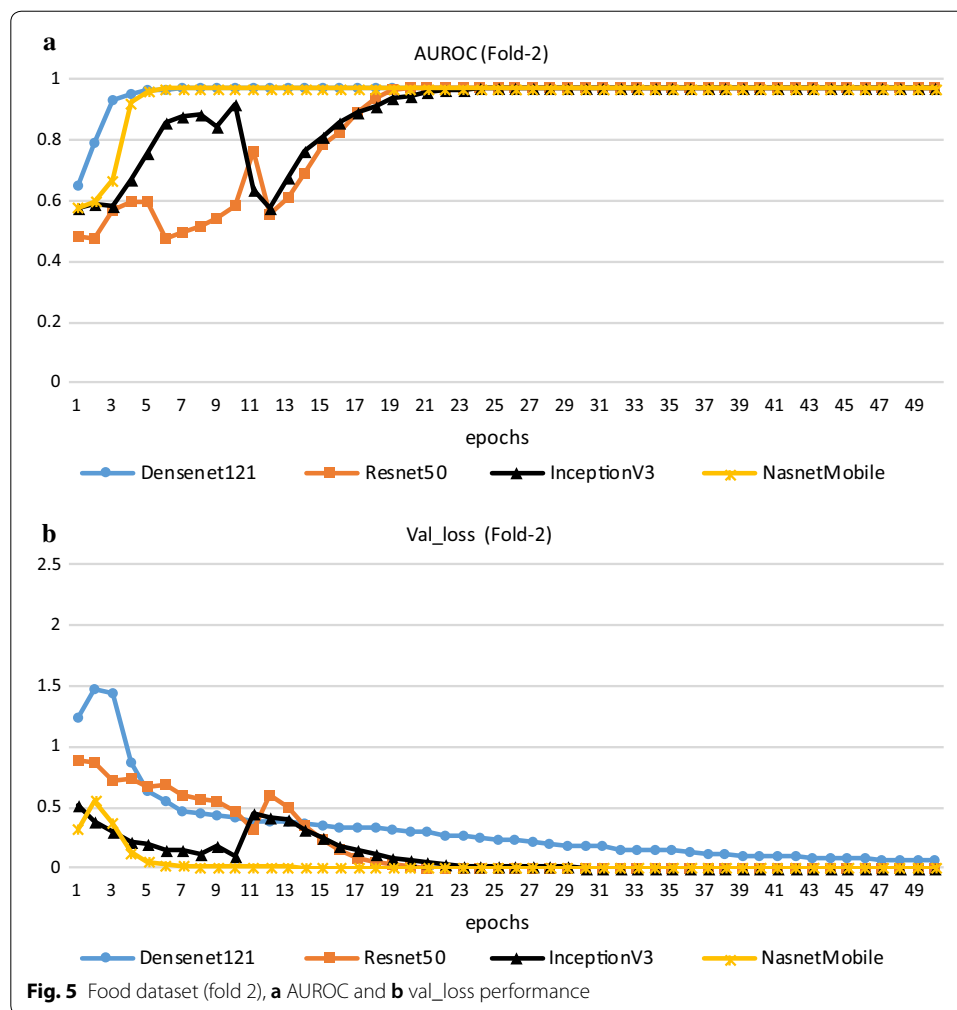


To enable a more extensive evaluation, a threefold dataset was employed. Therefore, the data of all the images were used as training, testing, or validation data. The traditional food image dataset included 1644 images for 34 types of food. Each food corresponded to 30–50 pictures.

Model analysis and evaluation

The dataset was evaluated using several metrics, including AUROC, precision, recall, F1 score, and training time. The CNN network models used to test the performance included Dense121, ResNet50, InceptionV3, and Nasnet Mobile. The computer specifications were the same for all the experiments; no other users used the computer during the simulations. The computer specifications were as follows: CPU: Dual 20 Core Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz; RAM-512 GB 2133 MHz DDR4 LRDIMM; Storage 4 × 1.92 TB SSD Raid 0; GPU 8 × Tesla V100 1 Petaflop GPU Memory 128 GB; Network: Dual 10 GbE 4 IB EDR; Software and OS: Ubuntu Linux Host OS; Power: 3200 W.

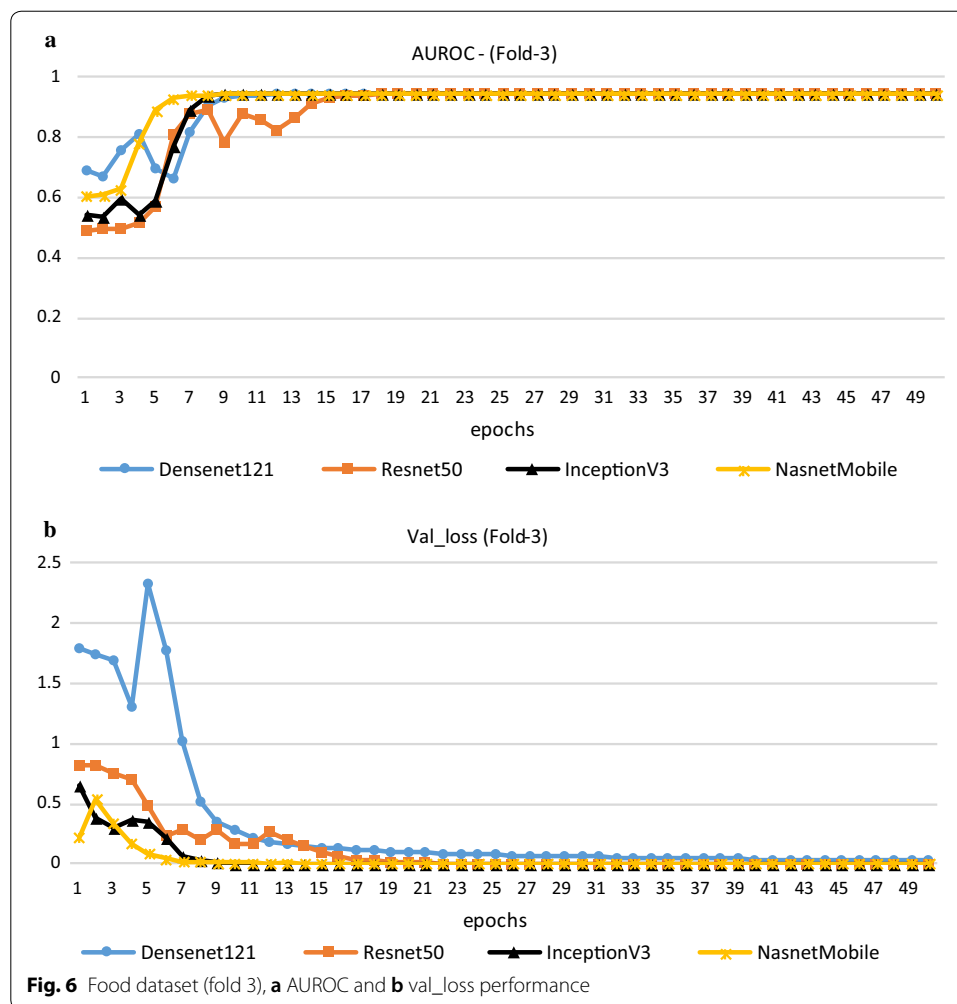
In the experiments, the metrics were obtained (AUROC, precision, recall, F1 score, and time execution) for each component of the testing data (fold 1, fold 2, and fold 3). The AUROC test was performed for each epoch during the training process. Figures 4



and 5 show the results for the training process evaluation. The test results for the testing data are presented in Fig. 6.

15th epoch. Moreover, the val_loss for all the network models decreases with the increase in the epochs. As shown in Fig. 4b, the val_loss is already minimal (0.1–0.5) in the InceptionV3 and NasnetMobile network models in epochs one to five. The val_loss value of the Resnet50 and Densenet121 models in the 15th epoch ranges between 0.5 and 1.7. At the 50th epoch, all the network models have an AUROC value of 0.88. This finding indicates that each network model can adequately classify the dataset validation at each epoch. The dataset validation consists of 10% of the dataset.

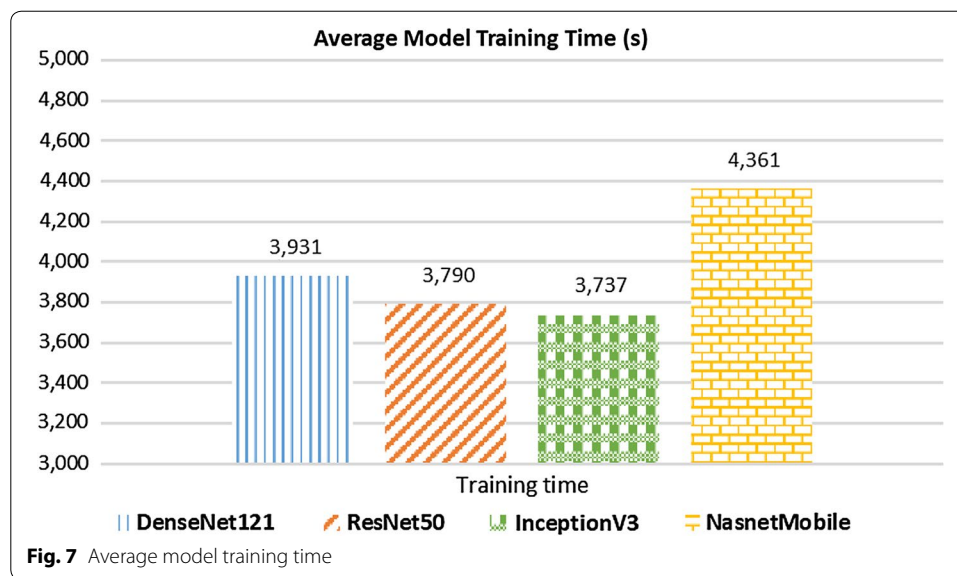
The fold 2 experiment has the same characteristics as the fold 1 experiment. As shown in Fig. 5a, all the network models exhibit a high AUROC value after the 20th epoch. This result is slightly different from that for fold 1 experiments, in which the AUROC value is stable in the 15th epoch. As shown in Fig. 5b, the val_loss (0.1–0.5) is minimal in the InceptionV3 and NasnetMobile network models in epochs 1–9. The corresponding values for the Resnet50 and Densenet121 range from 0.5–1.5. At the 50th epoch, all the network models have an AUROC value of 0.99. In fold 2 experiments, the AUROC value is lower than that in the fold 1 experiment (0.88).



According to Fig. 5a, of the four network models, Dense-net121 can attain an AUROC value of more than 0.9 at the 5th epoch, followed by NasnetMobile. The new Resnet50 and InceptionV3 models attain epoch values of more than 0.9 in the 19th epoch. Although Densenet121 has a reasonably large decrease in the val_loss compared to that in other network models, Densenet121 can rapidly achieve the maximum AUROC value in fewer epochs compared to the other network models.

From the results in Figs. 4 and 5, it can be concluded that Densenet121 undergoes slow learning owing to the gradual decrease in the val_loss. This aspect is advantageous as the model becomes generalized to the food dataset and rapidly attains a higher AUROC metric than the InceptionV3 and Resnet50 models. The literature also indicates that the Densenet121 can generate excellent results even with small training datasets.

The fold 3 experiment results exhibited the same trends as those for fold 1 and 2 experiments. As shown in Fig. 6a, all the network models exhibited a high AUROC value after the 17th epoch. As shown in Fig. 6b, the val_loss value (0.1–0.5) is minimal in the case of InceptionV3 and NasnetMobile in epochs 1 to 9, whereas the corresponding values for the Densenet121 and Resnet50 are 0.5–2.4 and 0.2–0.7, respectively. At the 50th



epoch, all the network models exhibit an AUROC value of 0.94, which is higher than that for fold 1.

According to Fig. 6a, among the four network models, NasnetMobile can achieve an AUROC value of more than 0.9 most rapidly, at the 5th epoch. This finding is in contrast to those of fold 1 and fold 2 experiments, in which the Densenet121 attains a value higher than 0.8 or 0.9 in fewer epochs.

As shown in Fig. 6a, the InceptionV3 has an epoch value of more than 0.9 in the 9th epoch, followed by Resnet50, which attains this value in the 17th epoch. From the results shown in Fig. 5, it can be concluded that NasnetMobile exhibits the highest performance when testing the fold 3 dataset. This result is different from that for fold 1 and fold 2 experiments, in which Densenet121 exhibited the best results.

The test results for the food dataset fold 1, fold 2, and fold 3 show that all the network models can achieve AUROC values of more than 0.8. Subsequently, testing was performed using the validation data. The AUROC results for each fold were as follows: the values for folds 1, 2, and 3 were 0.88, 0.97, and 0.94, respectively. In folds 1 and 2, a stable AUROC value was most rapidly achieved by Densenet121, at the 5th epoch for fold 1 and 5th epoch for fold 2. In fold 3, the most stable AUROC was most rapidly attained by Nasnetmobile, at the 5th epoch.

Figure 7 shows the average time for the fold 1, fold 2, and fold 3 scenarios for all the network models. InceptionV3 has the smallest simulation time of 3737 s, followed by Resnet50 (3790 s), Densenet121 (3931 s), and Nasnetmobile (4361 s), respectively. These results indicate that InceptionV3 and Resnet50 have a lower simulation time; however, the AUROC values for InceptionV3 and Resnet50 are not the highest. Densenet121 obtained the highest AUROC values in fold 1 and fold 2 scenarios, whereas Nasnetmobile exhibit the highest values in the fold 3 scenario.

The training time for Nasnetmobile and Densenet121 was 10 min and 5 min larger than that for InceptionV3. However, both the network models attained the highest AUROC value more rapidly than InceptionV3 and Resnet50. Thus, Densenet121

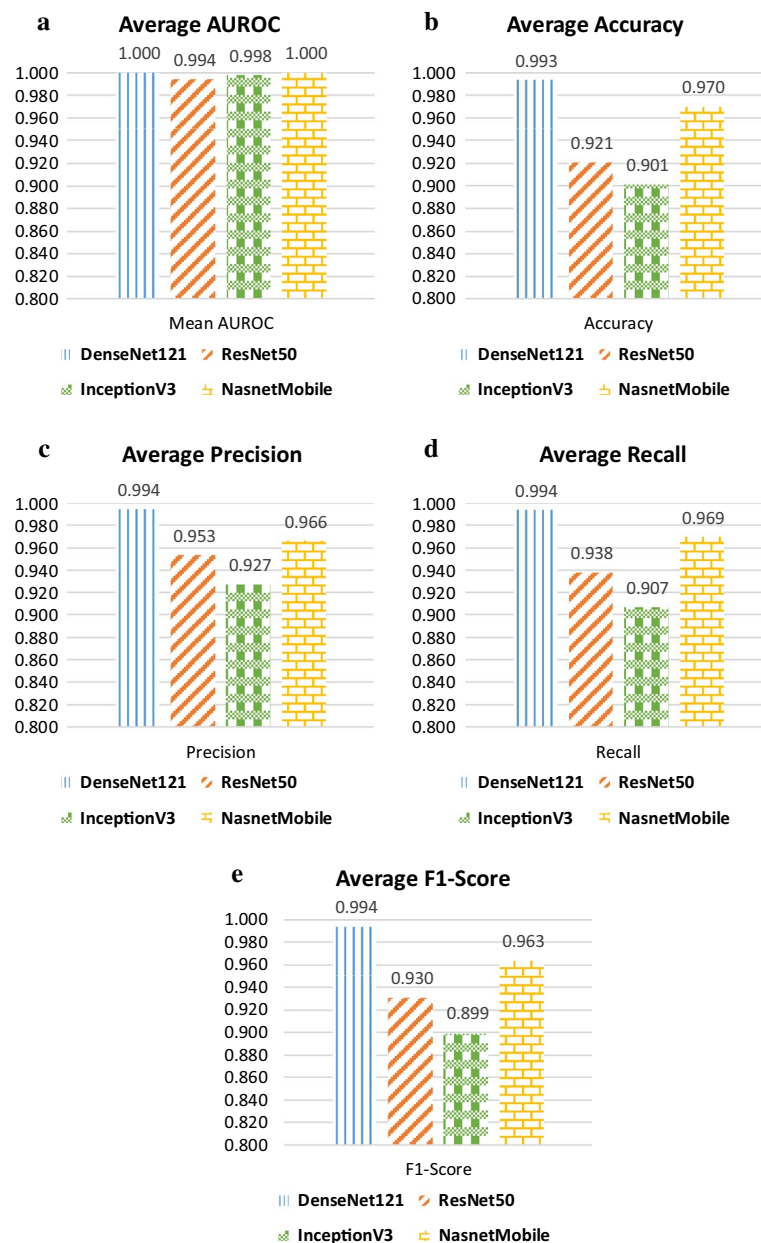


Fig. 8 Average testing performance **a** AUROC, **b** accuracy, **c** precision, **d** recall, **e** F1 score

requires a small epoch to obtain the maximum AUROC value. This aspect has significant implications in reducing the total time because a substantial repetition of epochs is not required in the training process to obtain the maximal AUROC value.

Experiments were performed on three folds, namely, folds 1, 2, and 3. The results for the dataset testing are shown in Fig. 8. The simulation results indicate that, in general, the evaluation of traditional food types using various network models produced satisfying results. The obtain AUROC values were more than 0.98, the accuracy values were higher than 0.90, precision and recall were higher than 0.92, and the F1 score was larger

than 0.88. These results could be obtained because the dataset was acquired professionally under the same ambient, exposure, and shooting settings for each image.

Due to the diverse ways of serving food items, the model encounters challenges when detecting the same food in a different dish. For example, Cendol can be served in a bowl or glass. If the training was performed using Cendol in a glass, the results for Cendol in a bowl may not be 100% accurate.

The measurement results show that the Densenet121 model attains the highest accuracy, precision, recall, and F1 score; all the values were more than 0.994. Furthermore, Nasnetmobile achieved values of more than 0.961 for all the metrics. Inception V3 achieved a high accuracy (0.901), precision (0.927), recall (0.907), and F1 score (0.899). Similar to Inception V3, Resnet50 also achieved a high accuracy (0.921), precision (0.953), recall (0.938), and F1 score (0.899). The high accuracy of Densenet121 occurs because of its architecture. As mentioned previously, the specific information that must be stored and the information that must be added are distinguished. This specific attribute of the Densenet successfully prevents the occurrence of the diminishing gradient phenomenon in the model.

The results indicated that the performance of Resnet50 and Inception V3 is not as high as that of Densenet121 and Nasnetmobile due to the occurrence of overfitting during training. Densenet121 and Nasnetmobile could maintain the value of val_loss at the end of the epoch as 0.081 and 0.0014 in fold 1, respectively. However, in fold 1, Resnet50 and InceptionV3 produced the minimum val_loss values, 0.00053 and 0.00033, respectively. This phenomenon was also observed in fold 2, in which Densenet121 and Nasnetmobile could maintain the value of val_loss at the end of the epoch as 0.061 and 0.0015, respectively. However, in fold 2, Resnet50 and InceptionV3 produced low values of the val_loss, that is, 0.00020, and 0.00097, respectively. The same phenomenon also occurred in fold 3. Densenet121 and Nasnetmobile could maintain the value of the val_loss at the end of the epoch as 0.025 and 0.0014, respectively, in fold 3. However, Resnet50 and InceptionV3 produced the minimum val_loss values of 0.00032 and 0.00033, respectively.

The value of the val_loss for Resnet50 and InceptionV3 is extremely low. Thus, these models have learning characteristics that are specific to the dataset being trained at each epoch iteration. This aspect is in contrast to Nasnetmobile and Densenet121, both of which can exhibit generalizations in each learning iteration. Consequently, no overfitting occurs, as indicated by the value of the val_loss being reasonably high. Moreover, the AUROC performance is enhanced with each iteration.

Conclusion

Traditional food knowledge is an essential aspect that must be preserved. In this research, professional data acquisition was performed to identify 34 types of traditional food from Indonesia. The images in the dataset were input to a model to predict the traditional food types in Indonesia. The evaluation results indicated that the proposed model exhibited excellent performance in predicting the type of traditional foods. In all the testing scenarios, the AUROC value was more than 0.98, accuracy was more than 0.90, precision and recall were more than 0.92, and F1 score was more

than 0.88. This result was achieved because the data were obtained professionally, and each image in the dataset was obtained under identical camera settings, exposure, and ambience. The deep learning network model with the highest performance is Densenet121 with an accuracy, precision, recall, and F1 score of 0.994, 0.994, 0.994, and 0.994, respectively. In the future, we aim to connect the proposed deep learning model with the supply chain of traditional food ingredients.

Abbreviations

TFK: Traditional Food Knowledge; AUROC: Area Under Receiver Operating Characteristic; UNWFP: United Nations World Food Program; MKL SVM: Multiple Kernel Learning Support Vector Machine; DPI: Depth Per Inch; CNN: Convolutional Neural Network; SVM: Support Vector Machine; COCO: Captioning Dataset Object Segmentation.

Acknowledgements

We would like to express our gratitude for the grant received from Universitas Indonesia (2020) PUTI Q1 No: NKB-1278/UN2.RST/HKP.05.00/2020 and to Tokopedia-UI AI Research Center for providing the infrastructure to conduct this study.

Authors' contributions

AW Conceptualization, preparation of introduction, methods, and analysis sections, investigation. HW Preprocessing of dataset, revision of introduction, methods, and analysis sections. PKF and ZPR Dataset preparation, data acquisition and performing simulations. PM and WJ Verification of experiment process, proofreading, critiquing. All authors read and approved the final manuscript.

Funding

Universitas Indonesia (2020) PUTI Q1 No:NKB-1278/UN2.RST/HKP.05.00/2020.

Availability of data and materials

<https://www.dropbox.com/sh/88stvg9krafl7z8/AAC97ldMbHe7ksp8Cc8l08K2a?dl=0>.

Competing interests

Not applicable.

Received: 31 August 2019 Accepted: 14 August 2020

Published online: 31 August 2020

References

1. Kwik JC. Traditional food knowledge: Renewing culture and restoring health (Master's thesis, University of Waterloo), 2008.
2. Porter JR, Xie L, Challinor AJ, Cochrane K, Howden SM, Iqbal MM, Travasso MI. Food security and food production systems, 2014.
3. Sharif MSM, Zahari MSM, Nor NM, Muhammad R. The Importance of knowledge transmission and its relation towards the malay traditional food practice continuity. *Procedia Soc Behav Sci*. 2016;222:567–77.
4. FAO, "FAO - News Article: Food diversity expresses cultural heritage and is key for healthy diets". Online. <http://www.fao.org/news/story/en/item/1171702/icode/>, 2019 Accessed: 2019.
5. Colozza D, Avendaño M, Urbanisation, dietary change and traditional food practices in Indonesia: A longitudinal analysis. *Soc Sci Med*. 2019.
6. WFP, Executive Brief: Indonesia Food Security Assessment and Classification. United Nations World Food Programme, 2007.
7. Limenta ME, Chandra S. Indonesian food security policy. *Indonesia Law Rev*. 2017;7:245.
8. Joutou T, Yanai K. A food image recognition system with multiple kernel learning. In 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 285–288, 2009.
9. Wazumi M, Han XH, Ai D, Chen YW. Auto-recognition of food images using SPIN feature for Food-Log system. In 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), pp. 874–877, 2011.
10. Ciocca G, Napoletano P, Schettini R. Food recognition: a new dataset, experiments, and results. *IEEE J Biomed Health Inform*. 2016;21(3):588–98.
11. Chen M, Dhingra K, Wu W, Yang L, Suktharankar R, Yang J, PFID: Pittsburgh fast-food image dataset. In 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 289–292, 2009.
12. Pouladzadeh P, Shirmohammadi S, Al-Maghrabi R. Measuring calorie and nutrition from food image. *IEEE Trans Instrum Meas*. 2014;63(8):1947–56.
13. FoodNet: Recognizing foods using ensemble of deep networks. *IEEE Signal Processing* Van Phat T, Tien DX, Pham Q, Pham N, & Nguyen BT. Vietnamese food recognition using convolutional neural networks. In 2017 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 124–129, October 2017.
14. Csáji B. Approximation with artificial neural networks. MSc. thesis 45 (2001).
15. Hnoohom N, Yuenyong S, Thai fast food image classification using deep learning. In 2018 International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI-NCON) pp. 116–119, February 2018.

16. Tiankaew U, Chunpongthong P, Mettanant V. A Food Photography App with Image Recognition for Thai Food. In 2018 Seventh ICT International Student Project Conference (ICT-ISPC), pp. 1–6. IEEE, July 2018.
17. Ege T, Yanai K. Estimating food calories for multiple-dish food photos. In 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR) (pp. 646–651), November 2017.
18. Zhang W, Zhao D, Gong W, Li Z, Lu Q, Yang S. Food image recognition with convolutional neural networks. In 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 pp. 690–693, August 2015.
19. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;34:1798–828.
20. Nguyen Q, Muckamala M, Hein M. Neural Networks Should Be Wide Enough to Learn Disconnected Decision Regions. *arXiv Prepr. arXiv1803.00094* (2018).
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ICLR.* 2015;75:398–406.
22. Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv Prepr. arXiv1602.07261v2* 131, 262–263 (2016).
23. Szegedy C. et al. Going Deeper with Convolutions. *arXiv:1409.4842* (2014). <https://doi.org/10.1109/cvpr.2015.7298594>.
24. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2818–2826 (IEEE, 2016). <https://doi.org/10.1109/cvpr.2016.308>.
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Multimed Tools Appl.* 2015;77:10437–53.
26. Krizhevsky A, Sutskever I, & Hinton, GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf. Process Syst.* 1–9 (2012).
27. Gu J, et al. Recent advances in convolutional neural networks. *Pattern Recogn.* 2018;77:344–77.
28. Lin M, Chen Q, Yan S. Network In Network. 1–10 (2013). <https://doi.org/10.1109/asru.2015.7404828>.
29. Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv Prepr. arXiv1602.07261v2* 131, 262–263 (2016).
30. Dong C, Loy CC, He K, Tang X. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell.* 2016;38:295–307.
31. Dauphin YN, Fan A, Auli M, Grangier D. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 933–941 (2017).
32. Pascanu R, Mikolov T, Bengio Y. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063 (2012).
33. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int J Uncertainty Fuzziness Knowledge Based Syst.* 1998;6:107–16.
34. Srivastava RK, Greff K, Schmidhuber J. Highway Networks. 2015. <https://doi.org/10.1002/esp.3417>.
35. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proc.-30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017* 2017–Janua, 2261–2269 (2017).
36. Kuen J, Kong X, Wang G, Tan YP. DelugeNets: Deep Networks with Efficient and Flexible Cross-Layer Information Inflows. *Proc.-2017. IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017* 2018–Janua, 958–966 (2018).
37. Larsson G, Maire M, Shakhnarovich G. Fractalnet: Ultra-deep neural networks without residuals. *arXiv Prepr. arXiv1605.07648* (2016).
38. Barret Z, Vijay V, Jonathon S, Quoc V. Le; Nasnetmobile, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8697–8710.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)