IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 15, 2021, accepted January 10, 2022, date of publication January 14, 2022, date of current version January 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143119

A Dish Recognition Framework Using Transfer Learning

TRUONG THANH TAI, DANG NGOC HOANG THANH, AND NGUYEN QUOC HUNG
College of Technology and Design, University of Economics Ho Chi Minh City, Ho Chi Minh City 72500, Vietnam

Corresponding author: Dang Ngoc Hoang Thanh (thanhdnh@ueh.edu.vn)

This work was funded by the University of Economics Ho Chi Minh City (UEH), Vietnam.

ABSTRACT Dish understanding from digital media is an interesting problem, but it also contains a big challenge. The challenge comes from the complexity of ingredients in the dish. With the development of deep learning, several effective tools can solve the problem partially. In this work, the task of dish recognition is considered. A novel dish recognition method based on EfficientNet architecture and transfer learning is proposed. First, we modify the EfficientNet-B0 by adding several important layers. Second, we use transfer learning to utilize optimal parameters obtained from pretraining the model on ImageNet, and then retrain it on a new dataset of dish images, i.e., UEH-VDR dataset. The UEH-VDR dataset contains images about Vietnamese dishes collected from various sources. Experimental results show that the proposed method can achieve an accuracy of 92.33% for the task of recognizing a dish. It also works more effectively than other models based on popular convolutional neural networks such as VGG and ResNet. In addition, a mobile application is also developed based on the trained data to serve visitors who want to discover the Vietnamese culinary culture.

INDEX TERMS Dish/food recognition/identification, transfer learning, deep learning, culinary tourism, EfficientNet, convolutional neural networks.

I. INTRODUCTION

Culinary tourism or food tourism [1] is a hot trend in the tourism map of the world. Dishes often contain the characteristics of each culture and local people. The characteristics of a culinary culture depend on many different factors such as geography, culture, religion, and climate of each country. For example, in Mediterranean cuisine, the most prevalent ingredients are olive oil and herbs. For Western European cuisine, steak and cutlets are the common dishes. In addition, sauces from cheese, grape wine are famous ingredients. For Eastern Asia and Indochina, rice is a primary food, and sauces are usually made from fish, shrimp, and soybean.

A large number of visitors are ready to pay for tasting characterized dishes of local regions [2]. Therefore, culinary tourism brings significant chances for economic development. Intending to improve the experience of culinary tourism, we focus on developing a system to recognize dishes. The system is developed based on deep learning with convolutional neural networks and transfer learning techniques. The system focuses on Vietnamese culinary culture with 9 famous traditional dish classifiers: Bun, Com-Tam, Pho, Goi-Cuon, Banh-Xeo, Banh-Trang, Banh-Tet, Banh-Mi, Banh-Chung. The system can be easily extended for a larger number of classifiers, as well as for dishes of other countries.

The significant contributions in this paper focus on the following manners. First, a novel convolutional neural network is developed based on EfficientNet and the transfer learning technique to recognize dishes. Second, providing a dataset of (Vietnamese) dishes and trained data with learned features to let researchers/scientists reuse and extend it with a larger number of dish classifiers. Finally, a mobile application is developed based on the trained data to solve the task of dish recognition, and it also provided some useful information about dishes.

The rest of the paper is organized as follows. Section 2 provided a systematic review of related methods. Section 3 presents materials and the proposed method to recognize dishes in 9 dish classifiers. Section 4 presents experimental results and the development of a mobile application to improve the experience of tourism visitors. Finally, Section 5 concludes the work.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

II. REVIEW OF RELATED METHODS

Dish recognition or food recognition is an interesting problem, and it has many applications in practice. It can be extended for analyzing calories in dishes to make an optimal meal in an aspect of calories for people. There are many works aiming at the problem. Chen *et al.* [3] proposed an automatic Chinese food identification and quality estimate based on traditional machine learning methods such as multi-label support vector machine (SVM) and Adaboost. Kawano and Yanai [4] developed a real-time mobile food recognition system with the help of SVM. The system can not only recognize dishes but also can estimate calories and nutrition. Yanai and Kawano [5] also developed another food image recognition system using a deep convolutional neural network (DCNN) with pretraining and fine-tuning. They developed the deep convolutional neural network based on Alexnet. Bolanos and Radeva used GoogleNet to develop a method for simultaneous food localization and recognition. Park *et al.* [6] proposed a new CNN architecture for Korean food image detection and recognition for mobile dietary management. Using VGG network, Yadav and Chand [7] developed another automated food image classification method. Based on residual learning, Martinel *et al.* [8] proposed wide-slice residual networks for food recognition. Horiguchi *et al.* [9] proposed a new CNN and it is known as a personalized classifier method for food identification. Zahisham developed another method for food recognition based on ResNet-50 model.

There are some primary drawbacks of the above methods. Because the dish recognition problem needs many resources and features to learn, the methods based on SVM or Adaboost have limited efficacy. Other methods based on CNN are more effective, but they require a lot of amounts of data for training. This is a characteristic of CNN models. In addition, some methods trained on complicated CNN models such as VGG, spent a lot of system resources.

In this paper, we also focus on CNN models due to their high efficacy. Moreover, we also notice the performance of CNN models that suits mobile applications. Therefore, we develop a dish recognition system using EfficientNet model and transfer learning technique.

III. MATERIALS AND METHODS

A. MATERIALS

A dataset of Vietnamese dish images named UEH-VDR, with 7848 images, was collected from multiple sources on the internet. All images are in RGB color mode with various sizes. UEH-VDR is published at the address *https://www.kaggle.com/truthtaicom/uehvdr-dataset*. The image dataset is split into 3 sets – training set with 6273 images (∼80% of random data), validation set with 780 images (∼10% of random data), and test set with 795 (∼10% of random data).

In Vietnamese culinary culture, there are many dishes, but they can be divided into several basic groups. For example, in the Pho group, there are several kinds of Pho such as Pho-Bo, Pho-Ga, Pho-Chay, and so on. To simplify, we just group them into a single name – Pho. And this name is also well-known by international visitors over the world. It is the same for the case of Banh-Mi – another famous Vietnamese dish. In the research, nine traditional and famous groups – Bun, Com-Tam, Pho, Goi-Cuon, Banh-Xeo, Banh-Trang, Banh-Tet, Banh-Mi, Banh-Chung, was used.

It is noted that the names of dishes are in Vietnamese and were not translated because we do not want to change their meaning as well as their popularity. Discovering a culture based on the names of local dishes is also an interesting experience for visitors. Figure 1 shows a part of samples of the UEH-VDR dataset.

The numbers of images of each class for the training set, the validation test, and the test set are demonstrated as in Figure 2.

FIGURE 1. A part of the Vietnamese dish image dataset.

B. METHODS

1) EfficientNet-B0

The family of efficient baseline networks (EfficientNet) [10] was developed based on convolutional neural networks by scaling all dimensions including width, height, and resolution, uniformly. EfficientNet-B0 is one of the architectures of the EfficientNet family and its architecture is presented in Figure 3. The EfficientNet-B0 architecture contains several layers:

Conv layer (Conv2D) is a 2D convolution layer. It creates a kernel convolved with the input of the previous layer to create the output.

Depthwise Conv layer will apply a single convolutional filter for each input channel.

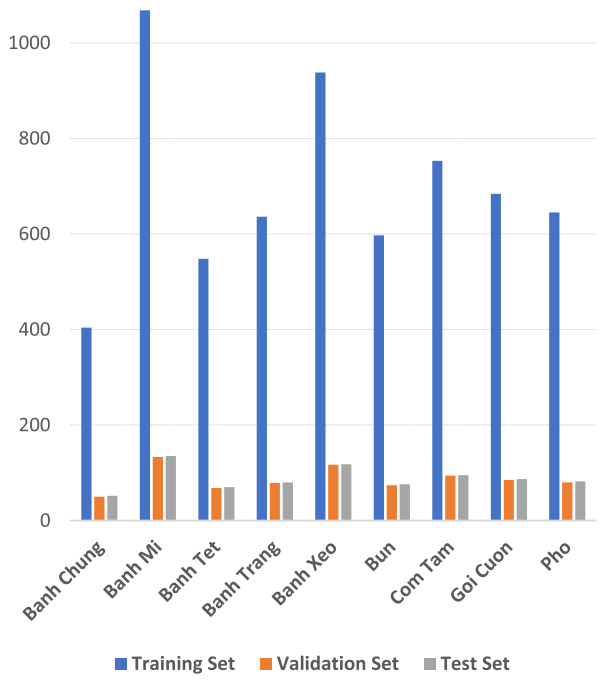

FIGURE 2. Number of images of each class in the training/validation/ test set.

BatchNormalization layer normalizes elements of the input by evaluating mean and variance of overall dimensions, and then it calculates the normalized activation function by:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \tag{1}$$

where x_i are elements of input, μ_B and σ_B^2 are mean and variance of overall dimensions, $\epsilon \ll 1$ is numerical stability constant.

Swish layer will apply a smooth and non-monotonic activation function for the input as follows:

$$swish\,(x) = \frac{x}{1 + \exp\,(-x)}. \tag{2}$$

SE layer will apply different weights for each channel instead of the same.

EfficientNet-B0 can perform well and stably on ImageNet, and it can be also transferred to other datasets effectively.

2) TRANSFER LEARNING FOR EfficientNet-B0

The transfer learning technique can be implemented based on any machine learning models, but it became more popular when using with deep learning. It is noticed that a convolutional neural network (CNN) will be trained on a given dataset to extract features. Based on the learned features, the model can predict results. To improve the accuracy of CNNs, a large amount of data is necessary to train the model. In practice, there are some difficulties to collect a large amount of data. And another case – if a part of the collected data is missing, it is not reliable enough to be implemented with CNN models. The solution to these issues is the transfer learning technique. Figure 4 presents an exploration of the transfer learning technique.

With transfer learning, a model can maintain optimal parameters based on the training process on a popular dataset such as ImageNet. Then, learned features are reused for the next learning task. Therefore, the overall accuracy of the model will be improved.

In this work, transfer learning will be applied for EfficientNet-B0 trained on ImageNet. The architecture of the

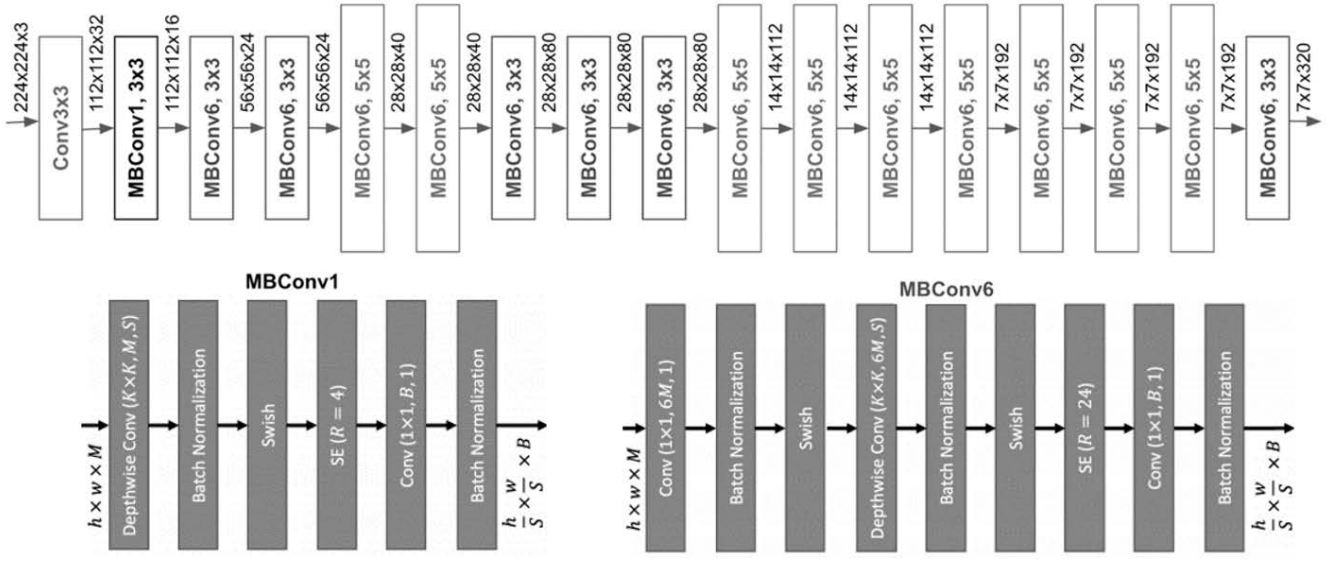

FIGURE 3. EfficientNet-B0 architecture, where $K \times K$ – a filter size, S – stride, B – feature maps, $h \times w$ – image size.

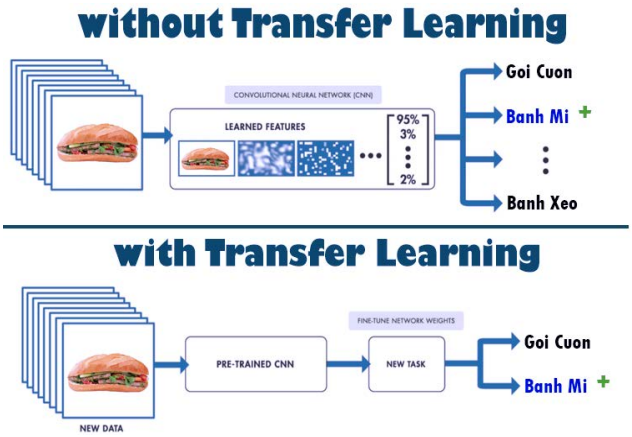

FIGURE 4. Transfer learning technique for a CNN model.

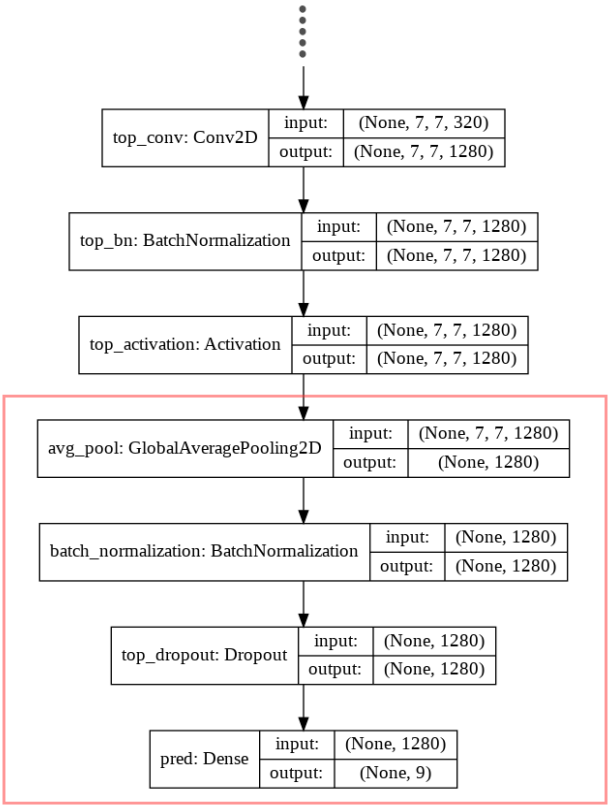

FIGURE 5. The proposed CNN using EfficientNet-B0 with transfer learning.

proposed model using EfficientNet-B0 with transfer learning is presented in Figure 5. Basically, the architecture of the proposed model is similar to one of EfficientNet-B0. However, at the end of architecture, we have added the following layers: GlobalAveragePooling2D, Dense, and BatchNormalization. In addition, to prevent the training process from overfitting, the dropout layer with the regularization technique was used. The goal of using GlobalAveragePooling2D layer is to reduce the size of the activations without sacrificing performance. Dense layer performs a deep connection,

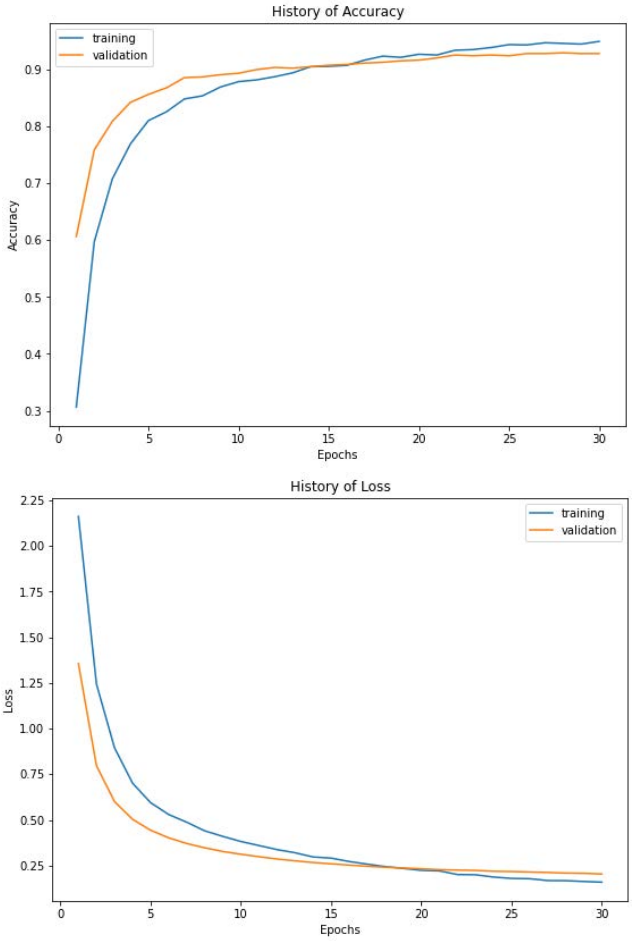

FIGURE 6. Training process of accuracy and loss.

i.e., it receives input from all neurons of its previous layer. BatchNormalization layer normalizes elements of the input by evaluating the mean and variance of overall dimensions, and then it calculates the normalized activation.

C. ERROR METRICS

To evaluate the performance of the dish classification and to compare the performance with other similar networks, the following metrics will be used:

$$accuracy = \frac{TP + TN}{FN + FP + TP + TN}, \tag{3}$$

$$precision = \frac{TP}{TP + FP}, \tag{4}$$

$$recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1 - score = \frac{TP}{2TP + FP + FN}, \tag{6}$$

where TP, TN, FP, FN denote true positive, true negative, false positive, and false negative, respectively.

Moreover, we also use the area under the ROC curve (AUC) metrics. AUC is evaluated by the area defined by the ROC curve, and vertical and horizontal axes.

TABLE 1. Performance comparison of CNN models.

	Accuracy	Precision	Recall	AUC	F1-Score	Kappa
VGG-16 [7]	76.23	82.88	69.43	95.76	75.56	72.92
ResNet-50 [14]	81.51	89.86	73.58	97.28	80.91	78.96
EfficientNet-B0 [10]	82.98	88.59	78.11	97.75	83.02	80.57
EfficientNet-B0 with Transfer Learning	**92.33**	**94.13**	**90.82**	**99.08**	**92.45**	**91.28**

	Banh xeo	Com Tam	Goi Cuon	Pho	Bun	Banh Trang	Banh Tet	Banh Mi	Banh Chung
Banh xeo	0.81	0.02	0.06	0.00	0.04	0.00	0.06	0.00	0.02
Com Tam	0.00	0.96	0.01	0.01	0.01	0.01	0.01	0.01	0.00
Goi Cuon	0.09	0.01	0.86	0.01	0.00	0.00	0.01	0.01	0.00
Pho	0.00	0.05	0.01	0.85	0.01	0.03	0.05	0.00	0.00
Bun	0.00	0.01	0.01	0.03	0.92	0.00	0.02	0.00	0.01
Banh Trang	0.00	0.00	0.00	0.01	0.01	0.92	0.04	0.00	0.01
Banh Tet	0.00	0.00	0.00	0.00	0.00	0.02	0.98	0.00	0.00
Banh Mi	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.97	0.00
Banh Chung	0.00	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.96

FIGURE 7. The confusion matrix.

The ROC curve is plotted at different threshold values for TPR and FPR:

$$TPR = \frac{TP}{TP + FN}, \tag{7}$$

$$FPR = \frac{FP}{FP + TN}. \tag{8}$$

Finally, Cohen's Kappa (Kappa score) [17] is also used. The Kappa score is measured by

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{9}$$

where p_0 is the empirical probability, and p_e is expected agreement.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. SETTINGS AND MODEL TRAINING

We train the proposed model on the platform of Google CODLAB with the training set, the validation set, and the test set with a ratio of ∼8:1:1 as in Section II. The number of epochs is set to 30. For the loss function, we use the categorical cross-entropy [12]:

$$Loss = -\sum_{c=1}^{M} y_{o,c} \ln\left(p_{o,c}\right), \tag{10}$$

where c – a class label; $p_{o,c}$ – predicted probability observation o is of class c; $y_{i,c}$ – binary indicator (0 and 1) and it receives 1 if class label c is the correct classification for observation o; and M is a number of classes.

To train the model, we use the optimal Adam method [13] with default parameters of the Keras framework.

The above settings are also applied for all the models considered in this paper: VGG-16, ResNet-50, and EfficientNet-B0.

The training graphs of the accuracy and the loss are presented in Figure 6.

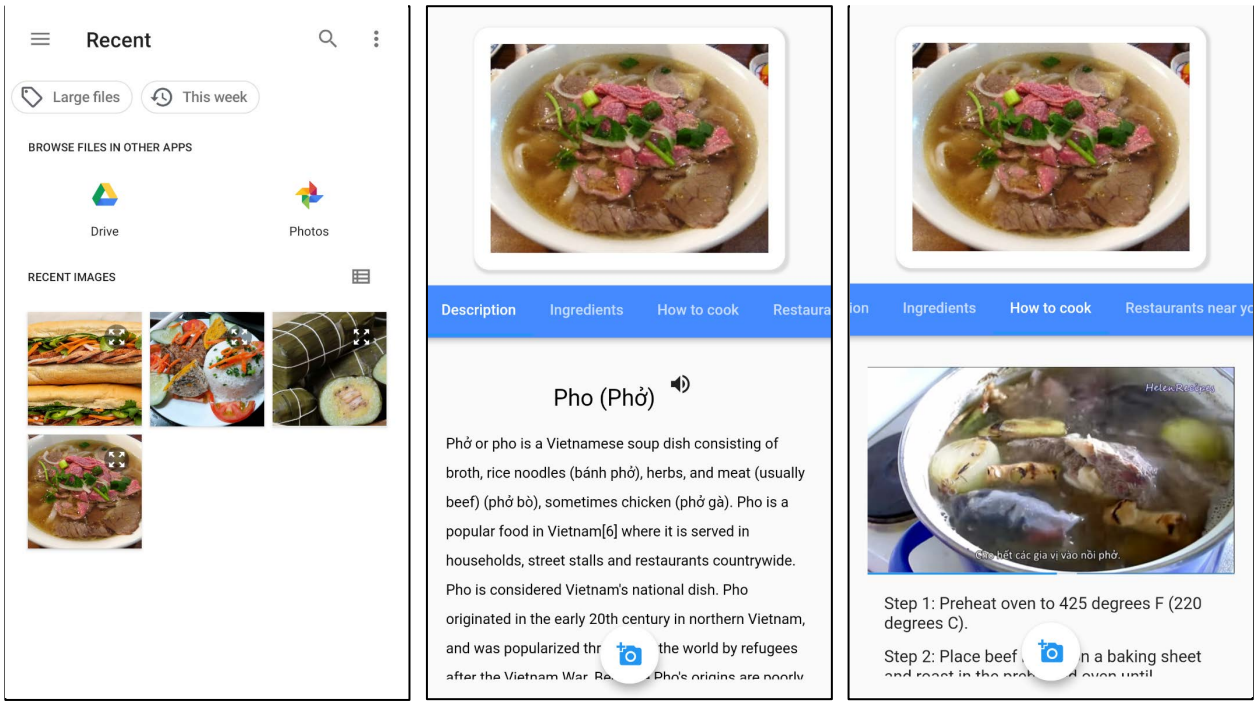

FIGURE 8. The user interface of UEH-VDR application: from left to right – Uploading an input image (first), Recognizing the dish and showing up its description (second), and showing up the how-to-cook (third).

B. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed method, we export the trained model and perform it on the test set. The confusion matrix with the input data is presented in Figure 7.

Based on the confusion matrix, we can evaluate the error metrics. Table 1 presents the values of accuracy, precision, recall, AUC, F1-score, and Kappa score of the proposed method along with VGG-16, Resnet-50, EfficientNet-B0. As can be seen that the EfficientNet-B0 with transfer learning achieved the best result for all metrics.

C. MOBILE APPLICATION DEVELOPMENT

To apply the results to Vietnamese dish recognition, we used the trained data to build a mobile application named Vietnamese Dishes Recognition (VDR). The application was developed based on the Flutter framework. Therefore, VDR app can work on multiple platforms including Android and iOS systems only with a single source code. The version of Android app is available on Google Play Store at the address: https://play.google.com/store/apps/details?id = com.ueh.vdr. We also have a plan to publish the app on Apple app store in the future.

Fig. 8 shows the user interface of the application. The application allows users to input an image by uploading it from the local system storage or capturing it by using the built-in camera. The application will recognize the dish classifier in 9 supported classifiers and show some useful information such as the name of the dish, and how to pronounce the name of the dish in Vietnamese, an introduction to the dish, ingredients, how to prepare it, and some restaurants to taste it.

V. CONCLUSION

In the article, we have proposed a novel method for dish recognition based on a combination of EfficientNet model with the transfer learning technique. The EfficientNet-B0 is modified by adding several important layers such as BatchNormalization, GlobalAveragePooling2D, Dropout, and Dense. Then, transfer learning was used to utilize learned features of the pretraining process on ImageNet. The proposed method achieved high accuracy and performance for dish identification. Moreover, a mobile application was developed to exploit the trained data to serve culinary tourism. Although the proposed method works effectively and stably, some limitations should be improved in the future such as increasing the number of dish classifiers, as well as combining with dish calorie evaluation.

ACKNOWLEDGMENT

This work was funded by the University of Economics Ho Chi Minh City (UEH), Vietnam.

REFERENCES

[1] B. Okumus, "Food tourism research: A perspective article," *Tourism Rev.*, vol. 76, no. 1, pp. 38–42, Feb. 2021.
[2] P. Liberato, T. Mendes, and D. Liberato, "Culinary tourism and food trends," in *Advances in Tourism, Technology and Smart Systems*. Springer, 2020, doi: 10.1007/978-981-15-2024-2_45.
[3] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung, "Automatic Chinese food identification and quantity estimation," in *Proc. SIGGRAPH Asia Tech. Briefs (SA)*, 2012, pp. 1–4.

[4] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 1–7.
[5] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jun. 2015, pp. 1–6.
[6] S. J. Park, A. Palvanov, C. H. Lee, N. Jeong, Y. I. Cho, and H. J. Lee, "The development of food image detection and recognition model of Korean food for mobile dietary management," *Nutrition Res. Pract.*, vol. 13, no. 6, pp. 521–528, 2019.
[7] S. Yadav and S. Chand, "Automated food image classification using deep learning approach," in *Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2021, pp. 542–545.
[8] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 567–576.
[9] S. Horiguchi, S. Amano, M. Ogawa, and K. Aizawa, "Personalized classifier for food image recognition," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2836–2848, Oct. 2018.
[10] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 6105–6114.
[11] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2018, pp. 1–11.
[12] D. P. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.
[13] Z. Zahisham, C. P. Lee, and K. M. Lim, "Food recognition with ResNet-50," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Eng. Technol. (IICAIET)*, Sep. 2020, pp. 1–5.
[14] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?" in *Proc. Brit. Mach. Vis. Conf.*, 2013, p. 5244.
[15] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–29, Dec. 2015.
[16] M. Bolanos and P. Radeva, "Simultaneous food localization and recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3140–3145.
[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, Apr. 1960.

TRUONG THANH TAI received the B.Sc. degree in business administration. He is currently pursuing the master's degree in information systems with the College of Technology and Design, University of Economics Ho Chi Minh City (UEH), Vietnam.

He was also qualified as a Programmer. He is also working with Tech JDI, Singapore. His research interests include machine learning, deep learning, business intelligence, image analysis, and knowledge discovery.

DANG NGOC HOANG THANH received the B.Sc. and M.Sc. degrees in applied mathematics, and the Ph.D. degree in computer science.

He is currently an Assistant Professor with the Department of Information Technology, College of Technology and Design, University of Economics Ho Chi Minh City (UEH), Vietnam. He has about 100 works on international peer-reviewed journals and conference proceedings, five book chapters, one book, and one European patent. His research interests include image processing, computer vision, machine learning, data classification, computational mathematics, fuzzy mathematics, and optimization. He is a member of scientific organization INSTICC, Portugal; ACM, USA; and IAENG, Taiwan. He is also a member of international conferences committee, such as the IEEE ICCE, Vietnam; IWBBIO, Spain; the IEEE ICIEV, USA; the IEEE ICEEE, Turkey; ICIEE, Japan; ICoCTA, Australia; and ICMTEL, U.K. He is also an Associate Editor of *The Journal of Engineering* (Wiley). He also served as a Guest Editor for *Current Medical Imaging* journal and *Frontiers in Applied Mathematics and Statistics*.

NGUYEN QUOC HUNG received the Ph.D. degree in computer science from the Hanoi University of Science and Technology (HUST), in 2016.

He is currently a Senior Lecturer and a Researcher with the Department of Information Technology, College of Technology and Design, University of Economics Ho Chi Minh City (UEH), Vietnam. His research interests include methods for acquiring, processing, analyzing in order to understand images, and human–robot interaction, artificial intelligence, HPC, and clouds computing.