IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 November 2022, accepted 2 December 2022, date of publication 12 December 2022,
date of current version 19 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3228701

RESEARCH ARTICLE

Food State Recognition Using Deep Learning

SAEED S. ALAHMARI[1], (Member, IEEE), AND TAWFIQ SALEM[2], (Member, IEEE)
[1]Department of Computer Science, Najran University, Najran, Saudi Arabia
[2]Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Saeed S. Alahmari (ssalahmari@nu.edu.sa)

This work was supported by the Deanship of Scientific Research, Najran University, under Award NU/-/SERC/10/589.

ABSTRACT Automated food detection and recognition methods have been studied to enhance end-user life. However, most existing research focused on food ingredient type recognition, with little work has been done for food ingredient state recognition. Successful recognition of food ingredient state plays a significant role in handling the food ingredient by an intelligent system. In this work, we propose a new novel cascaded multi-head approach based on deep learning to simultaneously recognize the state and type of food ingredients. We trained and evaluated the proposed approach on a benchmark dataset of food ingredient images with nine different food states and 18 food types. We compared the proposed approach with a non-cascaded deep learning approach. The cascaded approach shows improvement in food ingredient state recognition with 87% accuracy compared to 81% using a non-cascaded deep learning method. Our proposed method broadly applies to various tasks where food ingredient state recognition is essential, such as feeding elderly and disabled people and automating food recognition and preparation.

INDEX TERMS Food recognition, food state recognition, deep learning, features fusion, DenseNet.

I. INTRODUCTION

According to the U.S. Chamber of Commerce, the jobs requiring in-person attendance and having lower wages, including food service and hospitality, have suffered from labor shortages and had difficulty retaining workers [1]. Developing automated methods for food recognition and classification can help solve worker shortages by replacing human workers in many of the repetitive food preparation tasks in the industry. The computerized techniques for food preparation can be used in various applications, including supporting people with disability [2], especially people with vision impairment who need help recognizing the food type and ingredients. Our world has at least 2.2 billion people who are classified as having vision impairment, based on the World Health Organization (WHO) [3]. Therefore, it is essential to help this group of people in their daily life and their food recognition and classification needs. The current advances in the automation of services are induced by the effective learning approaches of artificial intelligence (AI), deep learning, and the availability of large data [4], [5], [6]. However, there are still challenges to fully automating services in interactive environments. Automation of food preparation requires the intelligent system to operate in an interactive kitchen environment while recognizing the food ingredient type and state. An example of the food ingredient type is orange and the state being sliced or whole. The food ingredient state is defined as the character of food which can be transformed by a human or robot chef interventions [7]. The same type of food can appear in different shapes and states depending on the intervention of a human or robot chef. For instance, a state of an orange fruit can be observed based on the texture and the shape of the fruit, which can be whole, peeled, sliced, or liquid (juiced).

Researchers have introduced different approaches for automation of food ingredient recognition [8], [9], [10]. These approaches use deep learning to learn representation from food images. However, these approaches learn to predict the food type or state independently, but the two tasks are related, and knowing the food type will help recognize the state. In this paper, we propose a novel approach to simultaneously learn to predict the food type and state in a cascading manner where the prediction of the food type will be fed into the prediction of the food state as those two are correlated. In our approach, food ingredient state recognition is achieved using the learned deep representations of food

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

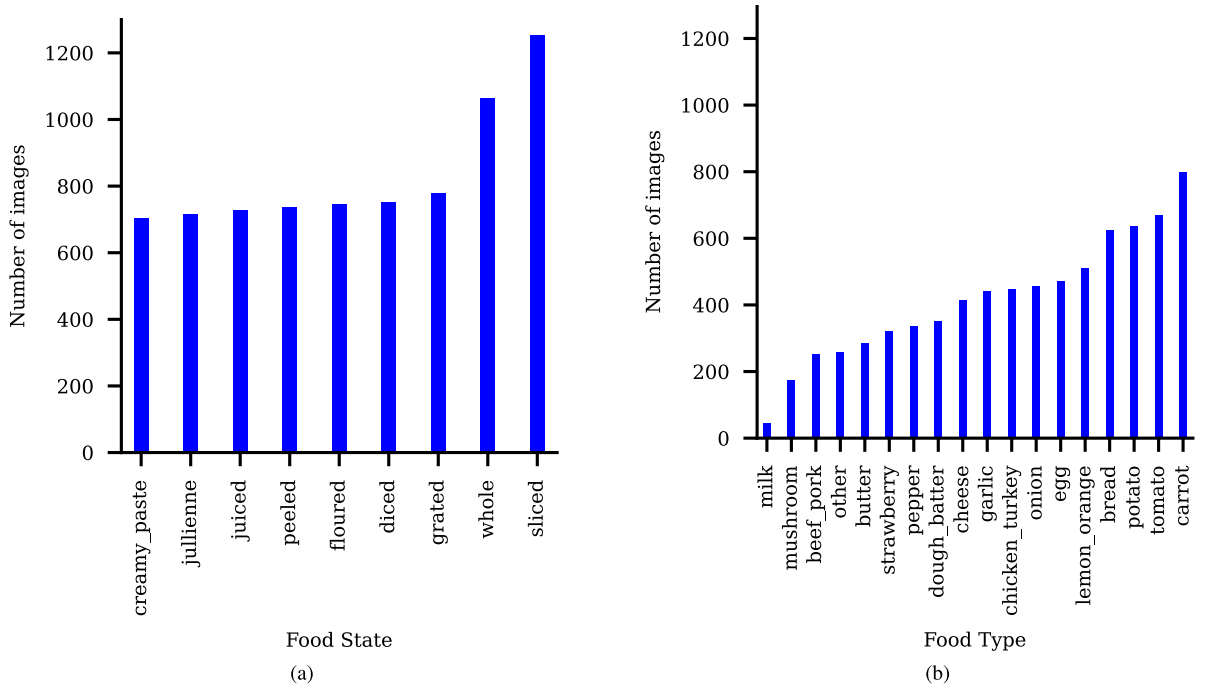


FIGURE 1. a) The number of images for each food state class, and b) the number of images for each food type class.

type and the input image's deep representations. To the best of our knowledge, our work is the first to address learning food ingredients type and state together in a cascading approach.

The Main Contributions of this work are:

- Proposing an approach, based on deep convolutional neural networks, to predict the food type and state in a multitasking and cascading manner.
- We manually labeled part of the dataset (test set) to use for the evaluation.
- Providing a detailed quantitative and qualitative evaluation of the learned models for various settings.

II. LITERATURE REVIEW

Different approaches have been proposed in the literature on food recognition. In [8], the authors proposed a deep learning-based approach for a food recognition system that allows the user to monitor the dietary intake during the day. This system uses a smartphone camera to take a picture of food as input to the trained deep-learning model and then provide the food classification and dietary information. This computer-aided food recognition system automates food recognition and dietary assessment to better monitor user dietary intake. A method for regression of food nutrition using deep learning is proposed in [11]. This approach uses Inception-v3 [12], ResNet [13], wide ResNet [14], and VGG16 [15] for learning food images to nutrition regression such as calories using ChinaMartFood-109 dataset [11]. Another method for food recognition using an ensemble of deep neural networks was proposed by Pandey et al. [9]. This approach assessed the use of traditional approaches (machine learning-based approaches) and an ensemble of deep neural network-based techniques for food recognition. The ensemble of neural networks produced the best result using the famous ETH Food-101 dataset [16]. In [17], the authors proposed a deep-learning approach to recognize traditional dishes with high calories. The proposed model was trained using EfficientNet pre-trained on ImageNet and fine-tuned using a dataset of traditional dishes images collected from the web. Then, the learned model was deployed on a smartphone for real-time inference.

Food recognition plays a significant role in helping impaired people. An approach for Middle Eastern food recognition was proposed in [18]. This approach used a pre-trained MobileNet-v2 and fine-tuning using a dataset of 23 classes [19], then deployed the learned model on phones for real-time inference.

In [20], the authors proposed a fusion of different pre-trained deep neural network-based classifiers for food recognition. This approach is based on fine-tuning several pre-trained deep learning models, then fusing the output predictions using a decision template. The authors have assessed this ensemble approach using two datasets, Food-11 and Food-101 [16], [21]. Salim et al. [22] studied different approaches for food recognition, including machine learning (traditional) and deep neural network-based approaches, where deep learning-based food recognition methods were the most effective compared to traditional approaches. DeepFood transfer learning approach was proposed in [23] for food type multi-class classification. The proposed approach extracts deep features from a pre-trained ResNet followed by feature selection and classification, where the results revealed improvement of food type multi-class classification using Mealcome (MLC) dataset [24]. A summary of food datasets and benchmark results and an evaluation of existing methods

for food recognition were presented in [25]. The authors trained the state-of-the-art method for five trials and achieved the state-of-the-art results on the UEC Food-100 dataset [26] by averaging the predictions of ResNeXt [27] and DenseNet models [28].

An improved VGG16-based approach was proposed in [29]. This approach used asymmetric convolution blocks instead of the original convolution kernel. Moreover, batch normalization was added to the VGG16, and a spatial attention mechanism was applied to improve the results of food type classification. To improve food recognition for vertical trait foods, a method was proposed by Martinel et al. [30] which used deep residual blocks and sliced convolution to learn recognition of vertical traits of food, such as a stack of pancakes. This approach improved the classification results using the Food-101 dataset.

Food recognition is important for automating the visual inspection of food quality and defects. A deep learning approach was proposed to detect defective apples and bananas in [31]. This approach uses multiple state-of-the-art deep learning architectures to recognize defective apples and bananas using food images. The deep learning architecture used in the work includes: ResNet-50 [13], DenseNet [28],MobileNet-v2 [32], NASNet [33], and EfficientNet [34]. The best performance was obtained using EfficientNet. Detecting the freshness of perishable fruits, including bananas, oranges, and apples, was studied in [35]. This approach applied transfer learning using AlexNet [36], VGG16 [15], and ResNet [13] architectures pre-train on ImageNet [37]. The dataset comprises six types of images: fresh banana, fresh orange, fresh apple, rotten banana, rotten orange, and rotten apple from an online dataset [38]. The best-performing model was obtained using ResNet architecture.

The previous approaches used static datasets, which represent a challenge because of the food appearance and shape variation. To solve this problem, a method that uses online continual learning was proposed for visual food classification [39] using the Food-1K dataset [40]. The approach first applied example selection using a similarity-based clustering approach for knowledge replay, and second, training online continual learning with a batch-based class balancing approach trained in a contrastive learning manner.

In [10], an approach to recognize the food state and type was proposed. This approach takes features extracted from an ImageNet-based pre-trained convolutional neural network (CNN), followed by a support vector machine for classifying food images into 20 food types and 11 food states. The authors experimented with multiple pre-trained CNN including GoogleNet [12], Inception-v3 [41], MobileNet-v2 [32], and ResNet-50 [13]. This approach independently learned food ingredient type and food ingredient state recognition using separate neural networks. The authors also proposed a new dataset of food ingredient images with state and type labels. However, this data was not made publicly available. Another approach for identifying the food ingredient state was proposed by Jelodar et al. [7]. The proposed approach used ImageNet pre-trained ResNet for fine-tuning using a dataset created by the authors. The proposed approach focused on learning food ingredient states only by fine-tuning pre-trained deep learning models [7]. Although these approaches can be applied to food ingredient state recognition, they suffer from shortcomings related to the learning process where learning food ingredients' states and types are performed independently.

III. DATASET

The dataset we used to learn and evaluate the proposed approach has annotated images of various ingredients in a kitchen. This dataset has 17 different most common cooking ingredients collected from over 250 online cooking videos from the two popular datasets [42], [43]. The cooking ingredients include: chicken/turkey, beef/pork, tomato, onion, bread, pepper, cheese, strawberry, milk, potato, garlic, egg, carrot, butter, mushroom, orange, and cheese [7].

Each food ingredient was labeled with a state where the state describes the status of the ingredient during the cooking process. In the original dataset, there are eleven different food ingredient state classes. However, there are two food ingredient state classes that do not have the object labels associated with each image, which are *mixed* and *other*. Therefore, we have eliminated these two food ingredient state classes (*mixed* and *other*) from the dataset. Thus, in the revised dataset, there are nine different food ingredient states where each food ingredient image has a state label and type label. An example of the food ingredient type is Orange, and the food ingredient state is sliced. The food state labels include whole, peeled, sliced, floured, grated, julienne, diced, juiced, and creamy paste. Our training set has 5251 images, the validation set has 1132 images, and the test set has 1180 images. The test set had only the state label. Therefore, we have manually labeled the test set for the food type, e.g., Orange or Potato. During training, we applied data augmentation, including rotation of 90°, 180°, 270°, horizontal and vertical flipping, and zooming. Figure. 1 shows the total images for each state and type label in the dataset. Table 1 illustrates the total number of images per food state and type classes on the test set. More information on the dataset collection and the labeling process is provided in [7]. Figure 2 presents examples from the dataset of different food ingredient types and states.

IV. RESEARCH METHODOLOGY

We propose a CNN architecture as shown in Figure 3 to estimate the two probabilities $P(f_t|img)$ and $P(f_s|img, f_t)$ for a food ingredient image *img*, where f_t represent the food ingredient type, f_s is the food ingredient state, and *img* represent the input image. For a given food ingredient image *img*, this approach learns two functions: the first is for estimating the probabilities of the food ingredient type using features learned from input images. The second function estimates the probabilities of the food ingredient state, which takes inputs from the learned representation of the food ingredient type in

FIGURE 2. Examples from the dataset, where each row corresponds to a food state class. The presented images are: creamy (first row), diced (second row), flour (third row), juiced (fourth row), and whole (last row).

concatenation with the food ingredient image feature vector. The proposed cascaded multi-head neural network integrates features learned for food ingredient type with the image-based deep features to learn to recognize food ingredient state effectively (cascaded approach). In other words, the proposed approach learns food state recognition with the guidance of food-type learned representations. Furthermore, the proposed cascaded multi-head neural network can simultaneously predict food state and type.

The input for our model is RGB images *img* of food ingredients. For this approach, we use the DenseNet121 model pre-trained on the ImageNet dataset and fine-tuning on the food dataset described in Section III. The earlier layers of DenseNet121 are set as non-trainable, whereas layers from the convolution layer (*conv*5) onward are set to trainable. Two heads on the top of DenseNet121 were added, where the first head is used for predicting the food ingredient type, and the second head is used for predicting the food ingredient state. Each head comprises three fully connected layers. The output of the second dense layer on the food ingredient type head is concatenated with the flattened feature vector of the image representation, and the concatenated feature vector is the input for the second neural network head consisting of fully connected layers for food ingredient state recognition as shown in Figure. 3. This neural network is trained in an end-to-end approach for learning food ingredient state and type simultaneously.

For the purpose of studying the impact of learning the food ingredient state and food ingredient type jointly in a cascaded manner, we have trained two models: the first model uses

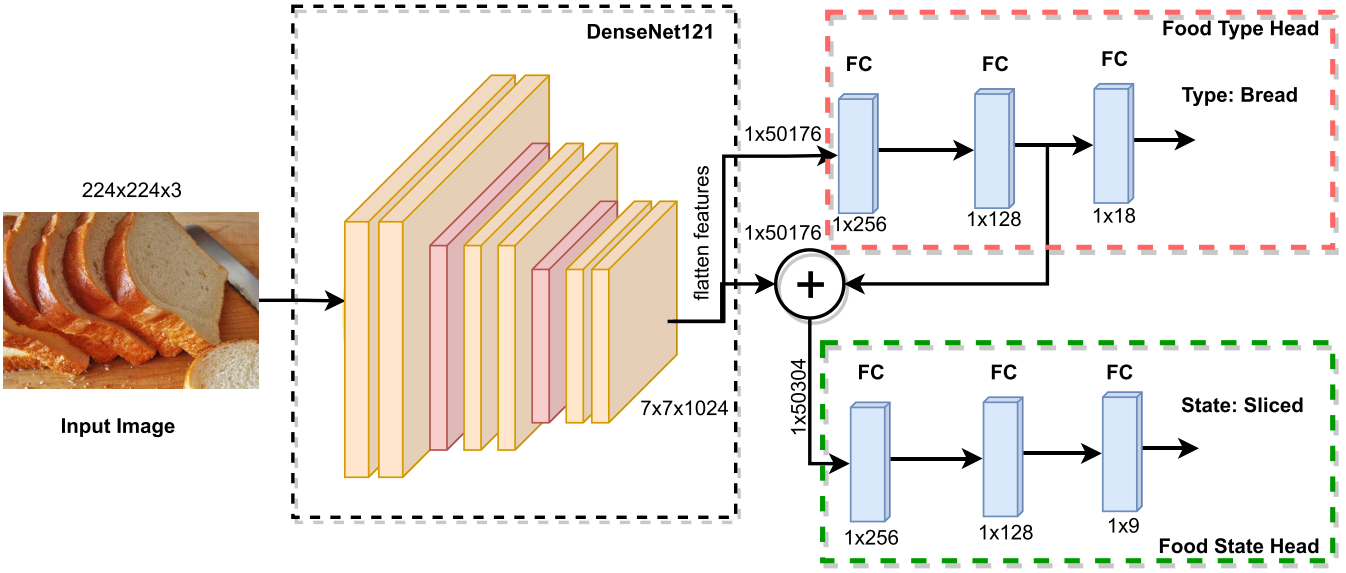


FIGURE 3. The proposed food ingredient state and type recognition cascaded multi-head deep neural network. This neural network learns food ingredient states and types simultaneously. The learned representation of the food ingredient type is concatenated with the image features to learn the food ingredient state.

TABLE 1. Total number of test set images per food type class is shown in the left table, and the total number of test set images per food state class is shown in the right table.

Food Type	Number of Test Images
Beef/Pork	49
Bread	105
Butter	43
Carrot	110
Cheese	72
Chicken/Turkey	81
Dough/Batter	65
Egg	75
Garlic	72
Lemon_orange	90
Milk	21
Mushroom	26
Onion	64
Other	26
Pepper	59
Potato	75
Strawberry	52
Tomato	95

Food State	Number of Test Images
Creamy/Paste	94
Diced	117
Floured	124
Grated	122
Juiced	133
Jullienne	131
Peeled	88
Sliced	196
Whole	175

DenseNet121 with a single head of three fully connected layers for learning to recognize the food ingredient state only. This model is called *non-cascaded single head model*. The second model has two heads on-top of DenseNet121 for the food ingredient type and food ingredient state outputs, as shown in Figure 3. This model concatenates the learned features from the last convolution layer and the second fully connected layer in the food type head. We called this model *cascaded multi-head model*, which learns to predict the food ingredient state and type, whereas the former model only learns to predict food ingredient states.

V. IMPLEMENTATION

Model architecture designing and coding was done using TensorFlow and Keras deep learning development libraries [44], [45]. Fine-tuning deep learning was performed on Nvidia GeForce 1080ti GPU architecture for 20 epochs. For the models' fine-tuning optimization, we used the Adam algorithm with a learning rate of 0.001, and the loss function was categorical cross-entropy. To ensure the repeatability of deep neural networks training, we have seeded all the libraries, and we set deterministic configurations using TensorFlow as described in [46].

VI. RESULTS

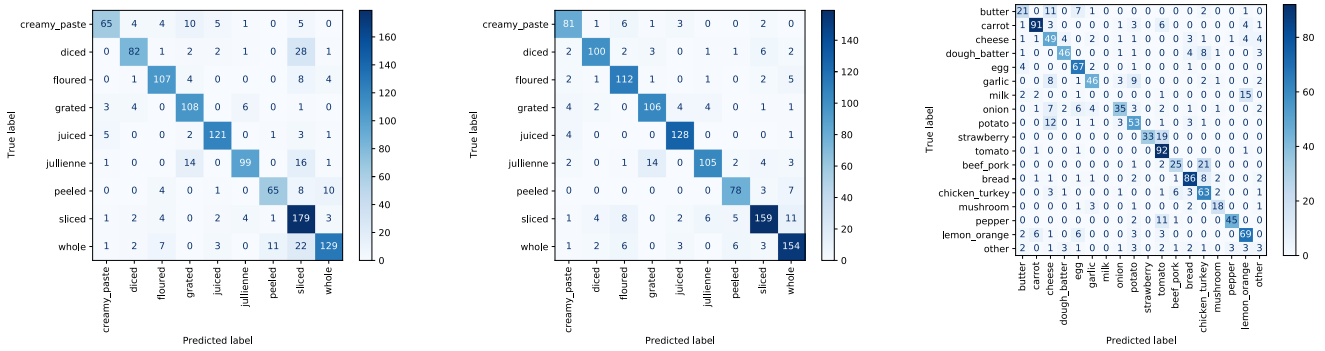
We have experimented with two classification approaches. First, we experimented with fine-tuning a deep learning model (DenseNet121) in an end-to-end approach where there is only one output head for the food state (i.e., non-cascaded single head approach). The second classification approach is for fine-tuning deep learning model (DenseNet121) in an end-to-end cascaded classification scenario where there are two neural network prediction heads: the first is for food ingredient type and the second is for the food ingredient state (i.e., cascaded multi-head approach). Furthermore, the latter approach uses fused learned food ingredient type representation with the image deep representation vector in a cascaded manner for learning to recognize the food ingredient state. Therefore, learning the food ingredient state is guided by the food ingredient type and image deep representations. In the following two subsections, we provided the results for our trained deep learning models.

A. NON-CASCADED SINGLE HEAD MODEL

This single-head model (i.e., the model trained to learn the food ingredient state only) is a non-cascaded approach where DenseNet121 is fine-tuned for food ingredient state recognition. The results of this approach showed an accuracy of about

TABLE 2. Results summary of two deep learning models. The second row shows results for the non-cascaded single-head model, where learning was done for food state only. The third row shows the results of learning from food state and type (i.e., the cascaded multi-head model). The results denoted by * are based on 11 state classes. However, our results use nine state classes, as discussed in the dataset section. The best result is in bold.

Model Name	Food State Results				Food Type Results			
	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score
Jelodar et al. [7]	80.40 *	–	–	–	–	–	–	–
Non-cascaded Single Head Model (Learning Food State Only)	80.90	0.82	0.81	0.81	–	–	–	–
Cascaded Multi-Head Model (Learning Food State and Type)	**86.69**	**0.87**	**0.87**	**0.87**	71.35	0.72	0.71	0.7



(a) (b) (c)

FIGURE 4. a) The confusion matrix for food ingredient state recognition (Non-cascaded single head model) for predicting the food ingredient state. b) The confusion matrix for food ingredient state and type recognition model (Cascaded multi-head model) where the two heads are for predicting the food ingredient state and food ingredient type. The confusion matrix is for the food ingredient state results. c) The confusion matrix for the food ingredient type results from the food ingredient state and type recognition (Cascaded multi-head model) approach with fusing feature vectors.

81%, a precision of 0.82, a recall of 0.81, and an F1-score of 0.81 for the food ingredient state.

B. CASCADED MULTI-HEAD MODEL

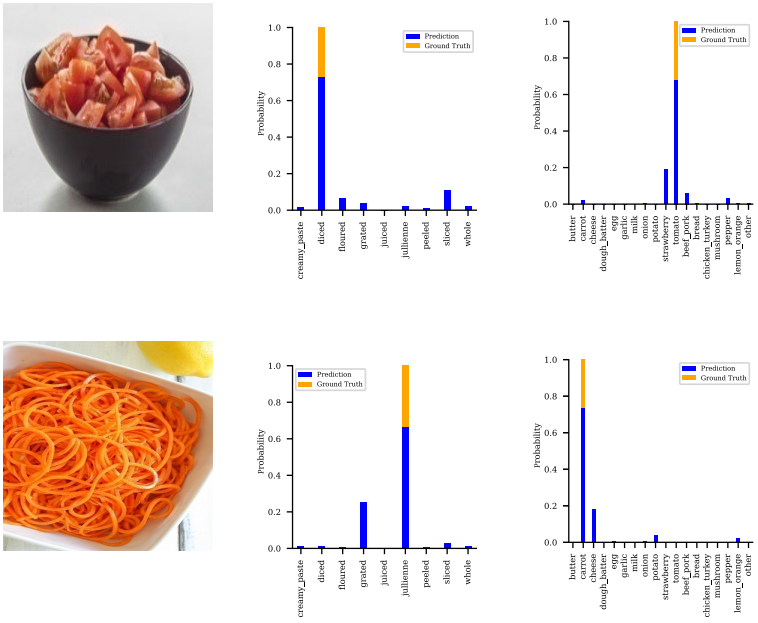
The food ingredient state and type recognition model is a cascaded multi-head model for classifying food ingredient type and food ingredient state simultaneously. This approach uses the food ingredient type feature vector and the input image feature vector to learn to recognize the food ingredient state. The two heads are built on top of DenseNet121, as shown in Figure. 3. This method showed superior results for food ingredient state classification compared to the former approach (i.e., non-cascaded single-head model), where accuracy, precision, recall, and F1-score are 87%, 0.87, 0.87, and 0.87, respectively. The food ingredient type accuracy is 71.35%, precision is 0.72, recall is 0.71, and F1-score is 0.70. The improvement of the food ingredient state results using the proposed model compared to the non-cascaded single head model shows that learning the food ingredient state using the representations learned for the food ingredient type in combination with the image features vector has a monumental results improvement. The results are shown in Table 2.

VII. DISCUSSION

The proposed approach aims to simultaneously learn food ingredient state and type in a cascaded manner. The cascaded multi-head deep neural network uses representations learned for food ingredient type recognition to learn food ingredient state. In other words, the learned deep features for food ingredient type are fused with the image deep representation for learning food ingredient state. Learning food state with fused representations (cascaded approach) shows superior results over learning food ingredient state without feature fusing *i.e., non-cascaded single head model*.

The food ingredient state recognition model (i.e., non-cascaded single head) for learning to recognize food ingredient states using the input image deep representations only showed that some food ingredient states are confused with the other food ingredient states. For instance, some diced labeled images are predicted as sliced, some creamy-paste labeled images cases are predicted as grated, some images labeled as whole food ingredient state is predicted as sliced, and some images labeled as julienne food ingredient state is predicted as sliced. This is because of learning the food ingredient state directly from the image representation without knowing the food ingredient type. The confusion matrix for a single-head neural network to predict the food state is shown in Figure 4a.

Fusing food ingredient type feature vector with image deep representations using the proposed *cascaded multi-head deep neural network* shown in Figure. 3 shows improvement in classifying the food ingredient state images. For instance, only three images of the test set were predicted as sliced, whereas the true label is whole. However, in the non-cascaded single-head neural network for learning to recognize the state of food ingredients, 22 images of the test set were predicted as sliced, whereas the true label is whole. Furthermore, the number of true positives for diced, creamy paste, floured, juiced, julienne, peeled, and whole increased using the *cascaded multi-head deep neural network* compared to using *non-cascaded single head deep neural network*. The



(a) Input Image (b) Food State (c) Food Type

FIGURE 5. Examples of the results of the proposed cascaded multi-head neural network, each row shows an image along with the food ingredient state head predictions and the food ingredient type head predictions.

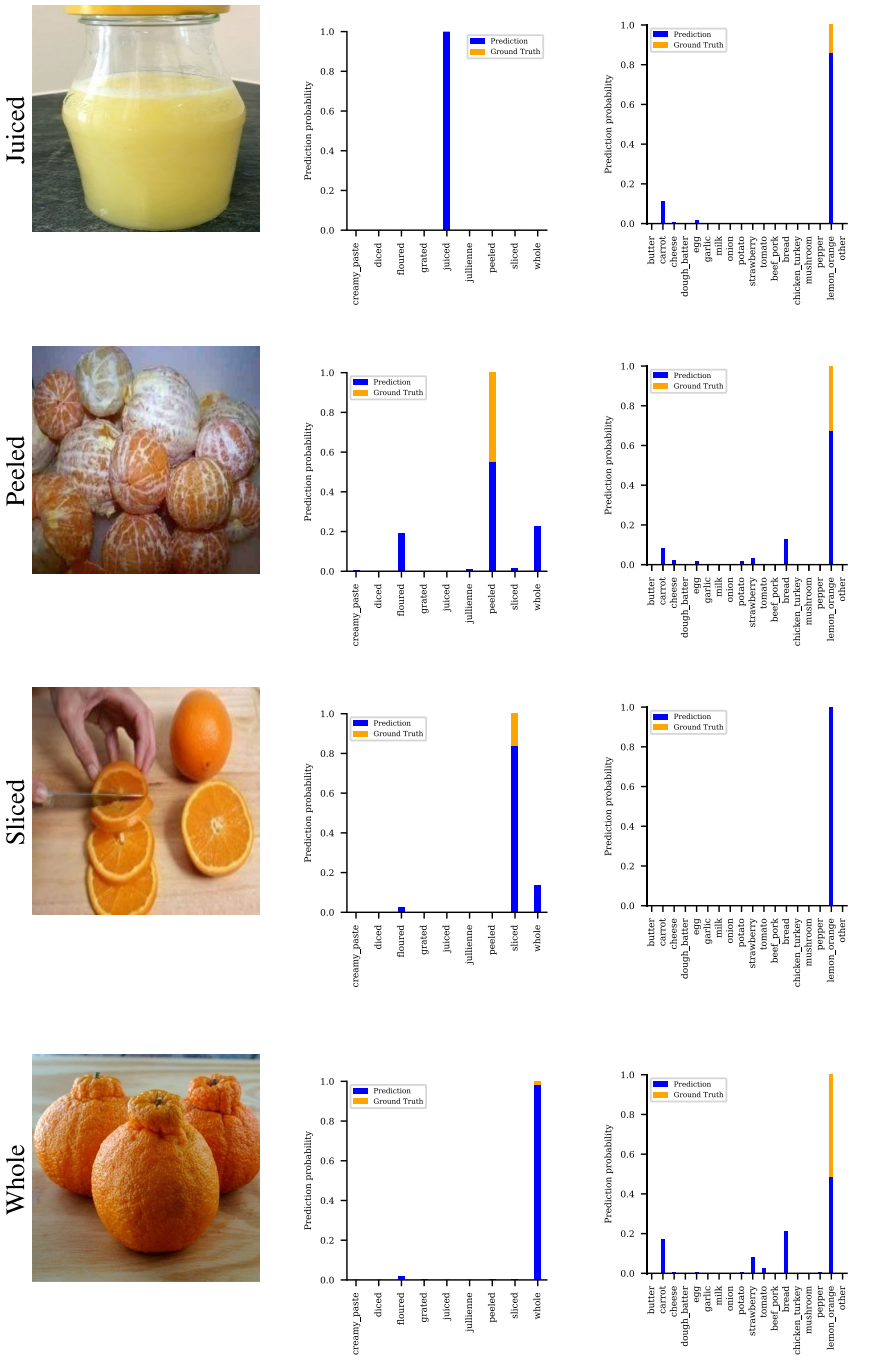
confusion matrix for the food ingredient state recognition using a cascaded multi-head deep neural network approach is shown in Figure. 4b, and Figure. 4c for the food ingredient type prediction. The food ingredient type head of the cascaded multi-head approach shown in Figure 3 showed prediction errors for some foods images that appear similar. For instance, milk and lemon_orange juice labeled images are similar; therefore 15 images of liquid milk were predicted as lemon_orange juice, as shown in Figure. 4c. Furthermore, the number of milk images in the dataset is low. Thus, the lower number of images labeled as milk could be the cause for the misclassification of milk images.

In Figure 5, two images are provided along with probabilities distribution for food ingredient state and type using the cascaded multi-head deep neural network. The ground truth label for each image is shown using the orange color. Figure 6 shows food state and type prediction results using the cascaded multi-head model for the same food type (orange fruit) but with different food states. In Figure 7, we provided visualization of the learned deep representations using the proposed *cascaded multi-head deep neural network* for both food state and type. This visualization was done by extracting deep features from the last convolutional layer of the fine-tuned DenseNet121, then applying the Grad-Cam internal representation visualization approach [47]. As observed from the deep representation visualization, our deep learning-based approach is focusing on the target objects presented in the images. Moreover, when there is more than one object of the same type in the image, the deep learning model focuses more on the closest object to the camera, as shown in Figure 7 top left image.

Convolutional Neural Network (CNN)-based food ingredient type and state recognition is an efficient approach for optimal results compared to handcrafted based approaches. Previous work showed that food discrimination could be done only using the food ingredient state images [10]. However, we found some challenges to food ingredient state recognition because of the similarity between food ingredients in terms of shape and texture, especially after manipulations. Furthermore, some states of food ingredients appear similar in images. For instance, julienne and grated food ingredient appearance are similar. Because of this apparent similarity, we noticed some cases with julienne food ingredients and predicted them as grated food ingredients.

Our work did not include deep feature extraction (transfer learning) or handcrafted feature extraction. Instead, this work focuses on learning the food ingredient state and type by fine-tuning an off-shelf neural network (DenseNet121) in an end-to-end manner. Some learned representations of the early layers of DenseNet121 were kept unmodified, where learning (fine-tuning) was done for the last layers of DenseNet121.

Although this project focuses on recognizing the state of food ingredients from a dataset collected from food preparation videos, this approach can apply to assist robots in other tasks, such as feeding elderly or handicapped people where an intelligent system needs to recognize the type and state of food ingredient for successful achievement of a certain task such as food preparation. Moreover, this research contributes to improving human-intelligent system interaction for better automation of services such as automation of feeding elderly persons and food ingredient grasping.



(a) Input Image (b) Food State (c) Food Type

FIGURE 6. **Examples of the results of the proposed cascaded multi-head deep learning approach, each row shows an image of an Orange fruit state along with the food ingredient state head predictions and the food ingredient type head predictions.**

The limitation and challenges of the proposed approach are related to the dataset. The dataset was collected by the authors of [7], where some state classes do not have the corresponding object label for food ingredient images. These classes are mixed and other state categories. Therefore, we had to remove the food state classes where dual labels for food state and type are not provided. Furthermore, the dataset we used suffers from data imbalance for some food type classes and poses a challenge for our proposed approach. Therefore, our future work focuses on solving the data imbalance issue for learning the food state and the type of food ingredient. Although the number of images in each food state and type class was low, we overcame this issue using a data augmentation approach.

Our future work includes addressing some shortcomings of the proposed approach, including improving the dataset by balancing the number of instances per class to improve

FIGURE 7. Grad-Cam visualization of the learned representation of the proposed cascaded multi-head neural network. The first and third-row visualize the latent space for the food state head. The second and fourth-row show visualization of the latent space of the food type head.

the results. Furthermore, we are planning to use the image segmentation of each food ingredient along with the raw food ingredient images for learning deep representation using deep learning.

VIII. CONCLUSION

Learning food ingredient states is an important task during food manipulation by an intelligent system. We propose an approach for learning food ingredient type and state jointly using a cascaded multi-head deep neural network. The learned feature vector for food ingredient type using deep learning is fused with image deep representation. The fused feature vector is used as input to the food ingredient state fully connected layers for the classification of food images. This approach showed superior results over the non-cascaded single-head neural network approach that learns to predict only the food ingredient state.

ACKNOWLEDGMENT

The authors would like to thank Ahmad Babaeian Jelodar from the University of South Florida for providing them with the dataset.

REFERENCES

[1] (Aug. 19, 2022). *U.S. Chamber of Commerce*. [Online]. Available: https://www.uschamber.com/workforce/understanding-americas-labor-shortage

[2] I. H. Sarker, "AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems," *Social Netw. Comput. Sci.*, vol. 3, no. 2, pp. 1–20, Mar. 2022.

[3] World Health Organization (WHO). *Blindness and Vision Impairment*, Accessed: Nov. 13, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[7] A. Babaeian Jelodar, M. Sirajus Salekin, and Y. Sun, "Identifying object states in cooking-related images," 2018, *arXiv:1805.06956*.
[8] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou, "A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure," *IEEE Trans. Services Comput.*, vol. 11, no. 2, pp. 249–261, Jan. 2018.
[9] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, "FoodNet: Recognizing foods using ensemble of deep networks," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1758–1762, Dec. 2017.
[10] G. Ciocca, G. Micali, and P. Napoletano, "State recognition of food images using deep features," *IEEE Access*, vol. 8, pp. 32003–32017, 2020.
[11] P. Ma, C. P. Lau, N. Yu, A. Li, and J. Sheng, "Application of deep learning for image-based Chinese market food nutrients estimation," *Food Chem.*, vol. 373, Mar. 2022, Art. no. 130994.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[14] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[16] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 446–461.
[17] N. U. Gilal, K. Al-Thelaya, J. Schneider, J. She, and M. Agus, "SlowDeepFood: A food computing framework for regional gastronomy," in *Proc. Smart Tools Apps Graph.—Eurograph. Italian Chapter Conf.* The Netherlands: The Eurographics Association, 2021, doi: 10.2312/stag.20211476.
[18] Ş. Aktı, M. Qaraqe, and H. K. Ekenel, "A mobile food recognition system for dietary assessment," 2022, *arXiv:2204.09432*.
[19] M. Qaraqe, M. Usman, K. Ahmad, A. Sohail, and A. Boyaci, "Automatic food recognition system for middle-eastern cuisines," *IET Image Process.*, vol. 14, no. 11, pp. 2469–2479, Sep. 2020.
[20] E. Aguilar, M. Bolaños, and P. Radeva, "Food recognition using fusion of classifiers based on CNNs," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2017, pp. 213–224.
[21] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained GoogLeNet model," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage.*, Oct. 2016, pp. 3–11.
[22] N. O. M. Salim, S. R. M. Zeebaree, M. A. M. Sadeeq, A. H. Radie, H. M. Shukur, and Z. N. Rashid, "Study for food recognition system using deep learning," *J. Phys., Conf.*, vol. 1963, no. 1, Jul. 2021, Art. no. 012014.
[23] L. Pan, S. Pouyanfar, H. Chen, J. Qin, and S.-C. Chen, "DeepFood: Automatic multi-class classification of food ingredients using deep learning," in *Proc. IEEE 3rd Int. Conf. Collaboration Internet Comput. (CIC)*, Oct. 2017, pp. 181–189.
[24] H. Chen, J. Xu, G. Xiao, Q. Wu, and S. Zhang, "Fast auto-clean CNN model for online prediction of food materials," *J. Parallel Distrib. Comput.*, vol. 117, pp. 218–227, Jul. 2018.
[25] B. Arslan, S. Memiş, E. B. Sönmez, and O. Z. Batur, "Fine-grained food classification methods on the UEC FOOD-100 database," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 238–243, Apr. 2022.
[26] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2012, pp. 25–30.
[27] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[29] H. Liu, H. Gong, and X. Ding, "Food image recognition algorithm base on improved VGG16," in *Proc. IEEE 2nd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, Dec. 2021, pp. 899–903.
[30] 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.
[31] N. Ismail and O. A. Malik, "Real-time visual inspection system for grading fruits using computer vision and deep learning techniques," *Inf. Process. Agricult.*, vol. 9, no. 1, pp. 24–37, Mar. 2022.
[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[33] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.
[34] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[35] A. Kazi and S. P. Panda, "Determining the freshness of fruits in the food industry by image classification using transfer learning," *Multimedia Tools Appl.*, vol. 81, no. 6, pp. 7611–7624, Mar. 2022.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2017.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[38] S. R. Kalluri. (Aug. 2018). *Fruits Fresh and Rotten for Classification*. [Online]. Available: https://doi.org/10.5281/zenodo.4788775
[39] J. He and F. Zhu, "Online continual learning for visual food classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2337–2346.
[40] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," 2021, *arXiv:2103.16107*.
[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[42] D. Paulius, Y. Huang, R. Milton, W. D. Buchanan, J. Sam, and Y. Sun, "Functional object-oriented network for manipulation learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2655–2662.
[43] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1194–1201.
[44] M. Abadi et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Accessed: Dec. 9, 2022. [Online]. Available: https://www.tensorflow.org/
[45] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io
[46] S. S. Alahmari, D. B. Goldgof, P. R. Mouton, and L. O. Hall, "Challenges for the repeatability of deep learning models," *IEEE Access*, vol. 8, pp. 211860–211868, 2020.
[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

SAEED S. ALAHMARI (Member, IEEE) received the Ph.D. degree in computer science from the University of South Florida, in 2020. He is an Assistant Professor of computer science with Najran University, Najran, Saudi Arabia. He has authored and coauthored many journals and conference papers. His research interests include learning from noisy and limited labeled data, machine learning, deep learning, medical image understanding, computer vision, and deep learning repeatability and explainability.

TAWFIQ SALEM (Member, IEEE) received the Ph.D. degree in computer science from the University of Kentucky, in 2019. He is a Visiting Assistant Professor with the Department of Computer and Information Technology, Purdue University, USA. His research interests include computer vision, remote sensing, medical imaging, and machine learning.