User-biased Food Recognition for Health Monitoring

Mazhar Hussain[1 [0000−0001−5054−6317]], Alessandro Ortis[1,2 [0000−0003−3461−4679]], Riccardo Polosa[2,3,4 [0000−0002−8450−5721]], and Sebastiano Battiato[1,2 [0000−0001−6127−2470]]

[1] Department of Mathematics and Computer Science, University of Catania, Viale A. Doria, 6, 95125 Catania, Italy.
[2] Center of Excellence for the Acceleration of HArm Reduction (CoEHAR), University of Catania, Catania, Italy
[3] Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy
[4] ECLAT Srl, Spin-off of the University of Catania, Catania, Italy
mazhar.hussain@phd.unict.it,{ortis,polosa,battiato}@unict.it

Abstract. This paper presents a user-biased food recognition system. The presented approach has been developed in the context of the FoodRec project, which aims to define an automatic framework for the monitoring of people's health and habits, during their smoke quitting program. The goal of food recognition is to extract and infer semantic information from the food images to classify diverse foods present in the image. We propose a novel Deep Convolutional Neural Network able to recognize food items of specific users and monitor their habits. It consists of a food branch to learn visual representation for the input food items and a user branch to take into account the specific user's eating habits. Furthermore, we introduce a new FoodRec-50 dataset with 2000 images and 50 food categories collected by the iOS and Android smartphone applications, taken by 164 users during their smoking cessation therapy. The information inferred from the users' eating habits is then exploited to track and monitor the dietary habits of people involved in a smoke quitting protocol. Experimental results show that the proposed food recognition method outperforms the baseline model results on the FoodRec-50 dataset. We also performed an ablation study which demonstrated that the proposed architecture is able to tune the prediction based on the users' eating habits.

Keywords: Dietary Monitoring · Food Recognition · Food Dataset · Artificial Intelligence for Health.

1 Introduction

Recognizing food from images is an extremely useful task for a variety of use cases. For example, it would allow people to track their food intake of what they consume by simply taking a picture, to increase awareness of their daily diet

by monitoring their eating habits, kind and amount of taken food, how much time the user spends eating during the day, how many and what times the user has a meal, analysis on user's habits changes, bad habits, and other inferences related to user's behavior and mood [1].It can help a doctor to have a better opinion with respect to the patient's behaviour, in the applications on quitting treatment response, smoke monitoring technology [2], dietary monitoring during smoke quitting [3] and smoking cessation system [4]. Food monitoring plays a vital role in human health that is directly affected by diet [5]. Humans life is strictly affected by the food, this encourages researchers to introduce new methods for food logging and automatic food dietary monitoring [6], food retrieval and classification [7]. This paper presents a novel food recognition method that takes into account the specific user to systematically analyse and infer his/her eating habits. The idea is to introduce a bias related to the user in the food classification pipeline. In particular, inspired by deep learning approaches applied on text representation learning [8], the proposed architecture learns a user's eating habits feature representation space. We also collected a new FoodRec-50 dataset that will be used for evaluation of the food recognition technology for dietary monitoring during smoke quitting. The rest of this paper is structured as follows. Section 2 describes the related works. Section 3 presents the details about the proposed food recognition method. Section 4 discusses the proposed method comparison and evaluation experiments. Finally, section 5 describes our conclusions.

2 Related Work

Recently, computer vision and deep learning techniques have gained a lot of attention due to high level of performances in various research fields and applications as well as in food recognition. Computer vision research devoted to the analysis of food images including previous works on food detection, classification, and segmentation. The paper in [9] covers food computing including acquisition, analysis, perception, recognition, retrieval, recommendation, prediction, and its applications in health, culture, agriculture, medicine, and biology. Different computer vision and machine learning techniques have been used for single-label food recognition, multi-label food recognition, food portion estimation, and personalized food recognition along with existing benchmark food datasets. The work in [10] presented a review on food recognition technology and its applications especially in the health department for dietary and calorific monitoring. Computer vision techniques for food understanding have been addressed in the areas such as food detection and recognition for automatic harvesting, food quality assessment for industry aims, dietary management, food logging and food intake monitoring, food retrieval, and classification with publicly available food datasets. Fakhrou et al. [11] proposed a smartphone application that utilizes a trained deep convolutional neural network (CNN) model for food and fruits recognition to assist children with visual impairments. Moreover, food recognition is improved using the ensemble learning approach with fusion of multiple

deep CNN architectures on a customized food dataset where soft voting method is used to ensemble multiple models results. A system [12] is proposed to effectively estimate nutrient intake by using RGB depth image pairs that are captured before and after meal consumption. This system consists of a novel multi-task contextual network for food item segmentation, classification with few-shot learning-based algorithms built by limited training samples for food recognition. Pfisterer et al. [13] developed an automatic semantic food segmentation method using multi-scale encoder-decoder network architecture for food intake tracking and estimation in long-term care homes. For the encoder, ResNet architecture trained on the imagenet dataset is used because of its discriminative feature learning ability. For the decoder, a pyramid scene parsing network is chosen. The proposed method achieved comparable results to semi-automatic graph cuts. The paper in [14] designed an automatic framework for tray food analysis to find the region of interest of the input image then predict the food class for each region. Different visual descriptors have been used including opponent gabor features, chromaticity moments, color histogram, local color contrast, gabor features, complex wavelet features, and convolutional features. A semisupervised generative adversarial network is used for food recognition [15] using partially labeled data. Network architecture consists of generator and discriminator. The generator produces dataset fake samples and discriminator learns the nature of the problem and further recognizes different food items with partially labeled training data. The author claimed outperformed results on the ETH Food-101 datasets and indian food dataset as compared to the AlexNet, GoogleNet, and Ensemble Net. The ResNet deep residual learning architecture [16] is proposed for image recognition with powerful representational capability for learning discriminative features from complex scenes. ResNet network architecture designed for the classification task, trained on the imageNet dataset of natural scenes that consists of 1000 classes. Evaluation has been performed with the residual network with depths of 18-layers, 34-layers, 50-layers, 101-layers, and 152-layers. The ResNeXt [17] architecture is a combination of ResNet and InceptionNet that contains a stack of residual blocks with a split-transform-merge structure in each block. This design introduced a new dimension that is cardinality or the size of the set of transformations and further Hu et al. [18] introduced a new architectural unit squeeze-and-excitation block that comprises the squeeze operation and excitation operation with the aim to enhance the quality of representations produced by a network. Related studies described above presents traditional food image recognition without taking into account the user habits. Our proposed study differs from the recognition that happens in the development of general purposes food recognition systems as proposed approach considers the specific user that uploaded the food image to learn and monitor its eating habits.

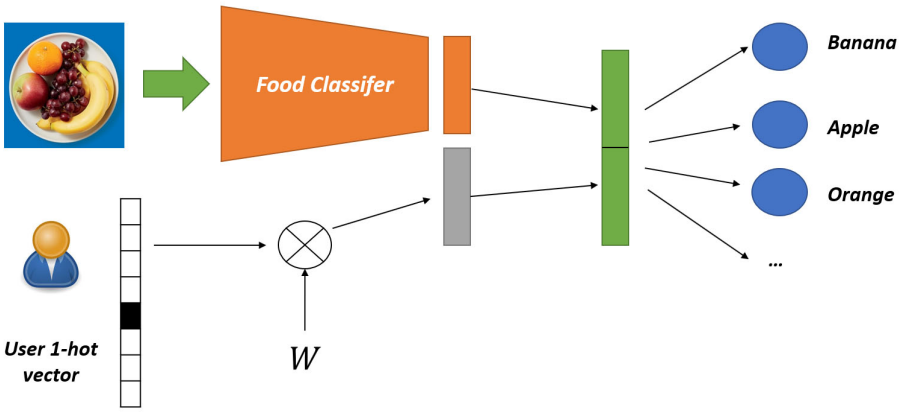

Fig. 1: Food recognition proposed architecture. Blue circles depict independent logistic activations for specific classes, which are activated by the presence of the food item in the visual content taking into account the bias given by the user.

3 Proposed Method

3.1 Data Acquisition and Annotation

The objective is to build a new unique robust dataset useful for the food recognition technologies development and evaluation stages. Our dataset is specific to the users who are involved in the smoke quitting process to monitor their dietary habits. The dataset is produced to study the correlation between eating information with smoking habits. In the future such data will be used to find correlations with respect to the smoking activity of the subjects during the period of observation. To collect the food data, the iOS and Android smartphone apps are used. Users can upload a meal intake image by taking a picture what they eat and can assign labels to the food image what it contains. Annotations are required during the supervised training of the network and also to test the examples during the evaluation phase for the food recognition method. To perform the experiments, we have first extracted food images for 50 classes from the FoodRec data. Although, some of the classes (beans, breadstick, carrot, chickpeas, corn, popcorn, grape, peas, zucchini, etc.) still have few images. Further, data is annotated manually for training and evaluation of the model which contains around 1100 images.

Table 1: Users and their eating frequencies

Food Items	Users Eating Frequencies					Food Items	Users Eating Frequencies				
	User 109	User 87	User 55	User 27	User 117		User 109	User 87	User 55	User 27	User 117
Almond	1	1	1	0	0	Green Tea	2	1	1	0	0
Apple	2	6	2	0	0	Jam	3	8	1	0	0
Arugula	3	1	4	0	0	Juice	3	1	2	0	0
Banana	1	3	1	0	0	Lentil	2	3	2	0	2
Bean	1	1	1	0	0	Meat	6	1	1	0	0
Biscuit	5	1	3	0	0	Milk	8	1	1	0	0
Blueberry	1	1	1	0	0	Mushroom	1	1	1	0	0
Bread	1	8	3	1	0	Orange	1	2	2	0	0
BreadStick	1	1	1	0	0	Pasta	6	1	3	1	0
Cake	2	3	2	0	0	Peas	1	3	1	0	0
Carrot	1	1	1	0	2	Pizza	5	2	2	0	1
Cereal	3	1	1	0	1	Popcorn	1	1	3	0	0
Cheese	2	2	2	0	0	Pork	2	1	1	1	0
Chicken	1	1	2	1	0	Potato	1	1	2	0	0
Chickpeas	1	1	2	0	0	Rice	2	2	1	0	0
Chips	2	1	1	0	0	Salad	1	2	1	0	0
Chocolate	1	2	2	1	0	Soup	2	1	1	0	0
Coffee	2	2	11	2	0	Spaghetti	2	2	1	0	0
Corn	1	1	1	2	0	Strawberry	1	1	1	0	0
Cracker	1	1	3	0	2	Tea	2	2	3	1	1
Croissant	2	1	1	0	0	Tomato	2	1	1	0	1
Doughnut	1	1	1	0	0	Tortellini	2	1	1	0	0
Egg	2	1	3	0	1	Vegetable	1	1	1	0	0
Fish	2	3	1	0	0	Yogurt	2	1	1	0	0
Grape	1	3	1	0	0	Zucchini	1	1	1	0	0

3.2 Proposed FoodRec Architecture

We proposed the FoodRec architecture for data coming from the FoodRec app [3], which is specific with respect to our purposes. The proposed system aims to recognize food items of specific users and monitor their habits. This task significantly differs from the recognition of any food instance depicted by a picture, such as happens in the development of general purposes food recognition systems. Figure 1 shows the proposed FoodRec architecture. In particular, a common multi-label food classifier is composed by a Convolutional Neural Network which defines a meaningful feature representation for the input images, based on the training task. Then, the representation is fed to multiple logistic units (i.e., blue circles in the Figure 1) which are activated if the associated food item is present in the picture. The proposed architecture will take into account the specific user that uploaded the picture. Indeed, since the proposed system is aimed to systematically analyse and infer user habits, our objective is to add to the food classification pipeline a bias related to the user. As consequence,

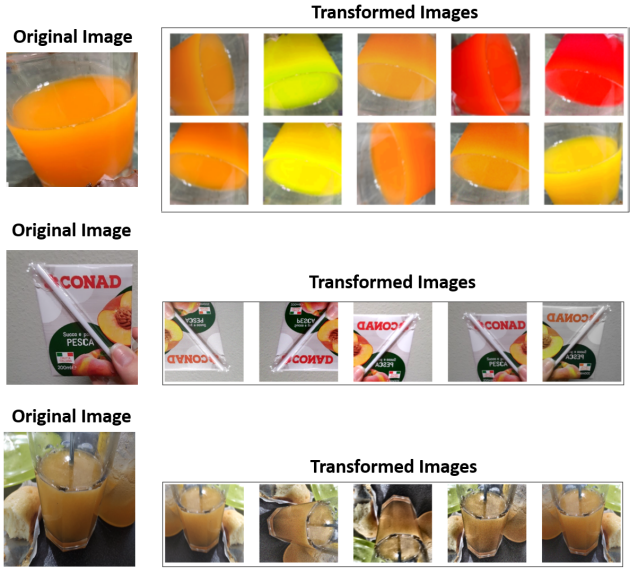

Fig. 2: Three different juices original images with transformed images.

the individual logistic activations will be fed with a feature that is obtained by concatenating the image and user feature. The latter one, is represented by the weight matrix W in Figure 1, which will be learned from the users' habits during the training stage.

3.3 User Data Annotation

As previously, we have extracted data for 50 classes but with only food items annotations. The proposed FoodRec architecture requires users annotation as well along with the food items eaten by them to train and test the model. So, data is annotated to embed the user information with image label so that we can feed the data simultaneously to the network. Now, the FoodRec images names contain the user ID and Image label. The idea is to extract the user ID from the image name while reading the images so that we implement the dataloader with the simultaneous user input and corresponding image data for both branches for the experiments.

3.4 Data Augmentation

Although FoodRec dataset consists of 1100 food images for 50 classes but some of the classes like beans, breadstick, carrot, chickpeas, corn, popcorn, grape, peas, zucchini, etc. still have very few images. Here we go with the data augmentation technique to deal with the lack of data. We have selected the top-20 users with the highest eating frequencies for all the food items. So, we have augmented

the data for these users to produce many altered and transformed versions of the same image. Image augmentation increases the training data as we don't have enough data with some food categories containing fewer food images and make a classifier more robust with a wide variety of transformed images. Different transformations are applied to the data such as image resize, image random crop, image horizontal and vertical flip, image random rotate, image motion blur, image optical distortion, image Gaussian noise, random brightness and contrast, CLAHE adaptive histogram equalization, hue, and saturation value. Three images have been taken from each food category and augmented for the top-20 users with the highest eating frequencies for all the food items. For example, Figure 2 shows three different juice images augmentation. Now, the FoodRec-50 dataset consists of around 2000 images after data augmentation.

4 Experimental Results

The baseline model consists of pretrained ResNet101 [16] model trained on ImageNet that is finetuned to extract a 1024-dimensional features vector to perform the food items classification. This model is trained using only the food images like the traditional classification algorithm without taking into account the user bias into the final decision making to classify the food items.

The proposed food recognition model consists of two branches, the food branch and the user branch to extract a 1024-dimensional concatenated feature map from these branches to recognize the food items. Food branch extracts 1024-dimensional feature map of food image using ResNet101 architecture with transferred weights from ImageNet dataset containing 1000 image categories, and further averaging pooling layer, flatten layer and fully connected layer are applied to get a 512-dimensional feature vector. The user branch extracts a 512-dimensional feature vector from the user bias using a fully connected layer. Finally, the output 1024-dimensional feature vector is obtained by concatenating the features extracted from both branches. User branch of the proposed network is learning 164 user eating weight vectors with 512 features. So, user weight matrix can be represented with 164 rows (one for each user) and 512 columns. The proposed network with one-hot user vector is effective because it learns specific user eating habits with the food image features as compared to the tradition food recognition systems. This approach is inspired by the document representation approach known as doc2vec presented by Le et al. [8]. Indeed, the model presented in [8] implements a document representation architecture in which the word/sentence features are affected by the document from which they have been extracted. In this way, the same word/sentence is represented differently depending on the source document which acts as a context for the encoded words and is represented as a one-hot input vector. The document branch is combined with another network branch devoted to represent single words, as we do combining the branch representing the food image with the one representing the user will act as a bias for the image representation.

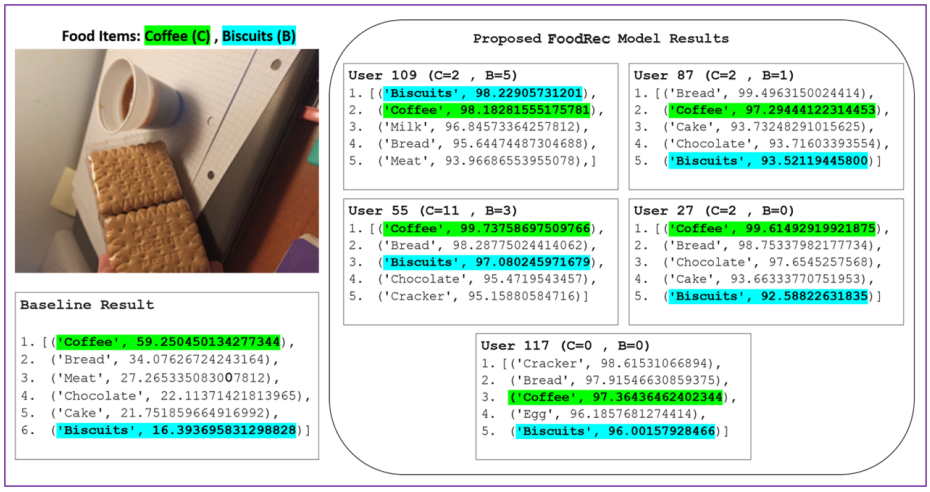

Fig. 3: Results comparison of test image containing Coffee and Biscuits.

The proposed and baseline networks are trained using same settings. Networks are trained using the Adam optimizer and the cross-entropy loss function. Initially, learning rate is set to 0.001 and it decays by a factor of 0.1 every 10 epochs. The batch size is set as 32 and networks are trained for 200 epochs. The FoodRec-50 data consists of 164 users eating different food items. Eating habits only for users 87, 109, 55, 27, and 117 for each food item is listed below in Table 1 with an individual food item and its eating frequency for that user. These users are chosen using the Euclidean Distance function to observe the difference in the decision making. The distance between user eating frequencies tells how much one user eating habits are different from other. Therefore, the distances between user eating frequencies have been used to select the users with different habits and perform specific tests aimed to assess the efficacy of our approach and its capability to encode the user eating habits. The effect can be observed in the Figures 3 and 4 showing the results. While the baseline method finds difficulties in the recognition of multiple food items in cluttered scenarios, the proposed method shows better performances, especially for users that have high frequencies for the food items present in the test image. We defined an "eating matrix" U x N, where U is the number of users (164 rows) and N is the number of considered food items (50 columns). Each row corresponds to the eating frequencies for each food item of a specific user. Then, we computed the distance for each user to find two users (87, 109) with a maximum distance between their eating vectors. Further, we calculated the sum of eating frequencies for each user and selected one user (55) with average and two users (27, 117) with the lowest sum of eating frequencies.

Table 2: Results comparison

Method	**User-biased**	**Top-5 Accuracy (%)**
Baseline Model	No	59.6
Proposed Model	Yes	71.1

4.1 Comparative Evaluation

The proposed FoodRec model results are compared with the baseline ResNet model. Figures 3 and 4 show multi-label food classification results comparison. Food and user concatenated representation fed to the logistics units containing sigmoid function to produce the independent probabilities for specific food classes. In particular, the output of each sigmoid is the probability that the input belongs to one specific food item. In other words, each sigmoid outputs $P(class = Banana|x), P(class = Bread|x)$, etc. Therefore, the score shown in Figures 3 and 4 for the food types is the percentage of the sigmoid output probabilities for each food item.

For example, Figure 3 contains a test food image with two food items coffee and biscuits with multi-label predictions. Baseline represents the ResNet model trained only with food images and results are shown below the test image in the figures. Then, the proposed FoodRec model results with five different users are shown next to the test image in the figures. We can observe from Figure 3 that the baseline model recognizes the coffee with a score 59.25 at the very first place but biscuits with a score 16.39 occur at the 6th place in the prediction order. On the other hand, the proposed FoodRec recognizes same two food items with improved score and occur in top five predictions for all five users. For the user 109, the top two predictions are biscuits and coffee with scores 98.22 and 98.18 respectively. This happens because the user with ID 109 has a relatively higher number of instances for these kinds of food in the eating matrix. For the user 117, although it did not drink coffee (C=0) or eat biscuits (B=0) but the model recognizes these food items at 3rd and 5th places respectively because the model learns both image and user features. Similarly, you can also observe the difference between the FoodRec and the baseline models in the Figure 4 with another test food image. The FoodRec model improves the score as well as learns the user dietary habits because the model is learning weight matrix for the users. As we change the user bias input, the result in the prediction is being changed according to the dietary habits for that user as you can obverse in the given figures. So, adding a bias related to the user to the food classification pipeline is effective to systematically analyse and infer user's habits. Users and their eating habits can be observed in Table 1. Moreover, the proposed model improves the general food recognition task with respect to the baseline model as shown in Table 2.

Fig. 4: Results comparison of test image containing Apple and Juice.

5 Conclusion

This paper aims to develop an automatic framework for food recognition using computer vision and deep learning techniques that plays a significant role to the health and food intake of people. The developed system acquires images of the food eaten by the user or subject over time which will then be processed by the proposed food recognition model to extract and infer semantic information from the food images. Experiments show that the proposed user-biased food recognition is effective and achieves higher results as compared to the baseline. The proposed model is hence able to influence the prediction by encoding the user bias as a result it also improves the task performance.

Acknowledgments

This investigator initiated study was sponsored by ECLAT srl, a spin-off of the University of Catania, with the help of a grant from the Foundation for a Smoke-Free World Inc., a US nonprofit 501(c)(3) private foundation with a mission to end smoking in this generation. The contents, selection, and presentation of facts, as well as any opinions expressed herein are the sole responsibility of the authors and under no circumstances shall be regarded as reflecting the positions of the Foundation for a Smoke-Free World, Inc. ECLAT srl. is a research based company from the University of Catania that delivers solutions to global health problems with special emphasis on harm minimization and technological innovation

References

1. Ortis, A., Farinella, G. M., Battiato, S. (2020). Survey on visual sentiment analysis. IET Image Processing, 14(8), 1440-1456.
2. Ortis, A., Caponnetto, P., Polosa, R., Urso, S., Battiato, S. (2020). A report on smoking detection and quitting technologies. International journal of environmental research and public health, 17(7), 2614.
3. Battiato, S., Caponnetto, P., Giudice, O., Hussain, M., Leotta, R., Ortis, A., Polosa, R. (2021). Food Recognition for Dietary Monitoring during Smoke Quitting. In IMPROVE (pp. 160-165).
4. Maguire, G., Chen, H., Schnall, R., Xu, W., Huang, M. C. (2021). Smoking Cessation System for Preemptive Smoking Detection. IEEE Internet of Things Journal.
5. Nishida, C., Uauy, R., Kumanyika, S., Shetty, P. (2004). The joint WHO/FAO expert consultation on diet, nutrition and the prevention of chronic diseases: process, product and policy implications. Public health nutrition, 7(1a), 245-250.
6. Kitamura, K., De Silva, C., Yamasaki, T., Aizawa, K. (2010, July). Image processing based approach to food balance analysis for personal food logging. In 2010 IEEE International Conference on Multimedia and Expo (pp. 625-630). IEEE.
7. Farinella, G. M., Allegra, D., Moltisanti, M., Stanco, F., Battiato, S. (2016). Retrieval and classification of food images. Computers in biology and medicine, 77, 23-39.
8. Le, Q., Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.
9. Min, W., Jiang, S., Liu, L., Rui, Y., Jain, R. (2019). A survey on food computing. ACM Computing Surveys (CSUR), 52(5), 1-36.
10. Allegra, D., Battiato, S., Ortis, A., Urso, S., Polosa, R. (2020). A review on food recognition technology for health applications. Health Psychology Research, 8(3).
11. Fakhrou, A., Kunhoth, J., Al Maadeed, S. (2021). Smartphone-based food recognition system using multiple deep CNN models. Multimedia Tools and Applications, 1-22.
12. Lu, Y., Stathopoulou, T., Vasiloglou, M. F., Christodoulidis, S., Stanga, Z., Mougiakakou, S. (2020). An artificial intelligence-based system to assess nutrient intake for hospitalised patients. IEEE transactions on multimedia, 23, 1136-1147.
13. Pfisterer, K. J., Amelard, R., Chung, A. G., Syrnyk, B., MacLean, A., Wong, A. (2019). Fully-automatic semantic segmentation for food intake tracking in long-term care homes. arXiv e-prints, arXiv-1910.
14. Ciocca, G., Napoletano, P., Schettini, R. (2016). Food recognition: a new dataset, experiments, and results. IEEE journal of biomedical and health informatics, 21(3), 588-598.
15. Mandal, B., Puhan, N. B., Verma, A. (2018). Deep convolutional generative adversarial network-based food recognition using partially labeled data. IEEE Sensors Letters, 3(2), 1-4.
16. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
17. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
18. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).