DEEPNOVA A DEEP LEARNING NOVA CLASSIFIER FOR FOOD IMAGES

SHADY ELBASSUONI , HALA GHATTAS , JALILA EL ATI , ZOULFIKAR SHMAYSSANI , SARAH KATERJI , YORGO ZOUGHBI , ALINE SEMAAN , CHRISTELLE AKL , HOUDA BEN GHARBIA , SONIA SASSI

SHADY ELBASSUONI , HALA GHATTAS , JALILA EL ATI , ZOULFIKAR SHMAYSSANI , SARAH KATERJI , YORGO ZOUGHBI , ALINE SEMAAN , CHRISTELLE AKL , HOUDA BEN GHARBIA , SONIA SASSI

IDRC GRANT / SUBVENTION DU CRDI : - TACKLING SCHOOL AND COMMUNITY DRIVERS OF CHILDREN'S UNHEALTHY DIETS IN ARAB CITIES

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 October 2022, accepted 1 December 2022, date of publication 8 December 2022,
date of current version 13 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227769

APPLIED RESEARCH

DeepNOVA: A Deep Learning NOVA Classifier for Food Images

SHADY ELBASSUONI[1], HALA GHATTAS[2,3], JALILA EL ATI[4], ZOULFIKAR SHMAYSSANI[1], SARAH KATERJI[1], YORGO ZOUGHBI[1], ALINE SEMAAN[5], CHRISTELLE AKL[2], HOUDA BEN GHARBIA[4], AND SONIA SASSI[4]

[1]Computer Science Department, American University of Beirut, Beirut 1107 2020, Lebanon
[2]Center for Research on Population and Health, American University of Beirut, Beirut 1107 2020, Lebanon
[3]Department of Health Promotion, Education, and Behavior, University of South Carolina, Columbia, SC 29208, USA
[4]Nutrition Surveillance and Epidemiology in Tunisia (SURVEN) Research Laboratory, National Institute of Nutrition and Food Technology (INNTA), Tunis 1007, Tunisia
[5]Department of Public Health, Institute of Tropical Medicine, 2000 Antwerp, Belgium

Corresponding author: Shady Elbassuoni (se58@aub.edu.lb)

This work was supported by the International Development Research Centre (IDRC) in Canada under Award 108641103657.

ABSTRACT Assessing the healthiness of food items in images has gained attention in both the computer vision and the nutrition fields. However, such task is generally a difficult one as food images are captured in various settings and thus are usually non-homogeneous. Moreover, assessing how healthy a food item is requires nutritional expertise and knowledge of the constituents of the food item and how it is processed. In this manuscript, we propose an end-to-end deep learning approach that can detect and localize various food items in a given food image using a customized object detection model. Our approach then assesses how healthy each detected food item is by classifying it into one or more of the four NOVA groups (Unprocessed Food, Processed Culinary Ingredients, Processed Food, and Ultra-processed Food). To train our food item detection model, we used two public datasets and a custom one we created ourselves and which contains images of food taken using wearable cameras. To train the NOVA food classifier, we use the custom dataset we created ourselves and that was manually labeled by expert nutritionists. Our food item detection model achieved a mAP of 0.90 and the NOVA food classifier achieved an average F1-score of 0.86 on test data.

INDEX TERMS Food images, NOVA food classification, deep learning, nutrition.

I. INTRODUCTION

With the proliferation of ubiquitous devices such as smart phones and wearable cameras, documenting dietary intake through images has become a common practice. Automatically assessing the healthiness of food items in images is a challenging computer vision task.[1] Food images can generally be non-homogenous as they can be taken in various settings and with different resolutions and qualities. They might also contain multiple food items in addition to other non-food related ones. Finally, assessing the healthiness of each food item in an image requires a knowledge of the constituents of the food items and how they are processed, which can be only accurately done by trained nutritionists who are familiar with the food items captured in the images.

To address all of the above mentioned challenges, we propose an end-to-end deep learning approach that can assess the healthiness of multiple food items in images without making any assumptions on how or where the images were taken. Our approach consists of two models: 1) a food item detection model that detects and localizes food items in an image that contains multiple food items, and 2) a classification model that classifies a detected food item into one or more of the four NOVA food groups, namely Unprocessed or Minimally Processed Food (Group 1); Processed Culinary Ingredients (Group 2); Processed Food (Group 3), and Ultra-processed Food (Group 4) [1]. Ultra-processed food

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

[1]In this manuscript, we use the term food to refer to both food and beverages.

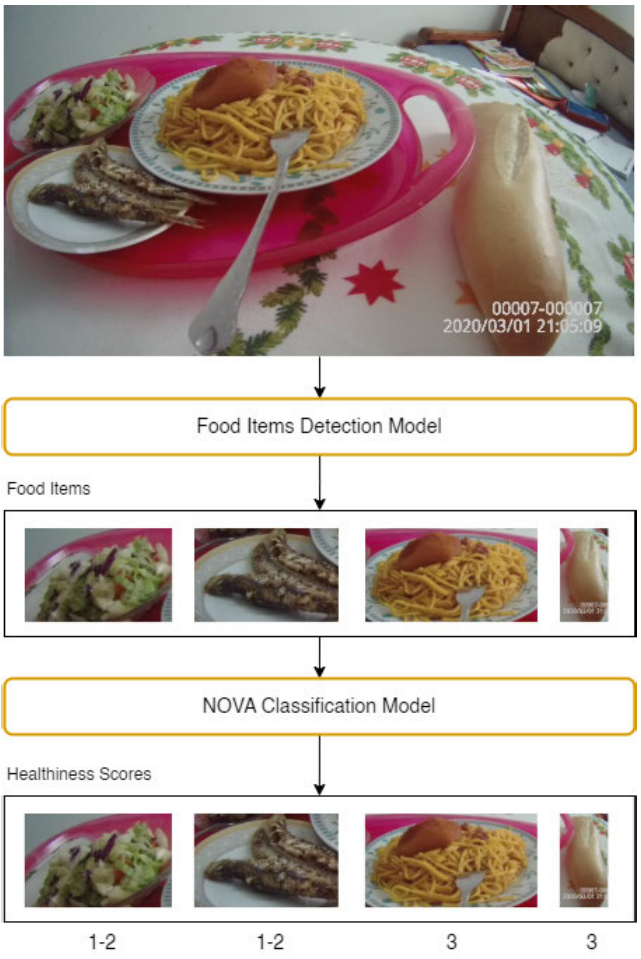

FIGURE 1. Overview of approach.

is usually considered unhealthy as it is associated with increased risk for obesity and other chronic diseases [2]. Using these two models, our approach can thus assess the healthiness of various food items in any given image. Our proposed approach is depicted in Figure 1. It takes as input an image containing one or more food items and passes it to the food detection model, which detects and localizes each food item in the images using bounding boxes. The localized food items are then extracted using their bounding boxes and each food item image is then passed to the NOVA classifier, which classifies the food item into one or more of the four NOVA groups based on the processing level it went through. In Figure 1, the predicted groups for each detected food item are indicated under the food item, where 1 indicates the first NOVA group, 2 the second one and so on.

Our food item detection model is based on a customized YOLOv3 (You Only Look Once, Version 3) model for general object detection [3]. To train our model, we used two public datasets, which are the EgocentricFood dataset [4] and the UECFood-256 dataset [5], and a custom dataset we created ourselves, which we refer to as the NOVA dataset. The first dataset, the EgocentricFood dataset, consists of food images captured using wearable cameras, and which contain food items of various types (general food, glass, cup, jar, can, mug, bottle, dish, and basket). The food items in the images in this dataset are localized using bounding boxes. The second dataset we used to train the food item detection model, the UECFood-256 dataset, consists of images of Asian food (e.g., Miso soup, Ramen Noodles and Fired Rice) that were crawled from the Web. The food items in each image in this dataset are also localized using bounding boxes. The third dataset we used to train our food item detection model, the NOVA dataset, consists of images of Tunisian food captured through wearable cameras. The images were then annotated on the crowdsourcing platform Labelbox[2] to detect food items in the images, again using bounding boxes. The reason we used these various datasets to train our food item detection model is to add more variability in the training data of the model to obtain a general model that can detect and localize food items in any food image no matter how or where it was captured. Our food item detection model achieved a mean Average Precision (mAP) of 0.90 on test data from the NOVA dataset.

Once food items have been detected and localized in a given food image, our approach extracts each food item using its bounding box and passes it to the NOVA classification model to estimate its healthiness. There have been many attempts to estimate the healthiness of food items using their corresponding images. However, most of these are quantitative approaches that are based on volume and calories estimation, which face many limitations. For instance, many of these approaches make unrealistic assumptions about the food images such as assuming the food images are all captured from specific predefined angles, or assuming the presence of reference objects in each image that can be used to estimate the volume of the food items. To this end, many expert nutritionists are advocating for food classification systems that are based on the food processing level rather than using calories and volume to assess the healthiness of food items [6]. Our proposed approach in this manuscript follows this school of thought by doing a qualitative assessment of the healthiness of food items rather than a quantitative one based on calories estimation.

Our NOVA classification model classifies food items into four groups according to the nature, extent and aim of the industrial processes that were applied to the food items. The first group is the Unprocessed or Minimally Processed Food, which includes natural food items such as vegetables, fruits, eggs, milk, water, etc. The second group is the Processed Culinary Ingredients, which is usually acquired from the first group and includes butter, oil, honey, etc. The third group is the Processed Food, which includes products made by adding salt, sugar or other Group 2 substances to Group 1 food such as unpackaged bread, canned fish, canned vegetables, etc. Finally, the fourth group is the Ultra-Processed Food, which includes food items that are produced using a series

[2]https://labelbox.com/

of industrial processes such as chips, chocolate, soft drinks, hotdog, etc. Obviously, a single food item might contain constituents that belong to more than one of these four groups and thus our NOVA classification model is a *multi-label* classification model.

Our NOVA classification model is based on the MobileNetV2 deep learning architecture [7]. To train our model, we used the NOVA dataset that we also used as part of the training data for the food item detection model. The dataset consists of image of Tunisian food that were captured using wearable cameras and that were annotated using crowdsourcing. In addition to localizing food items in the images using bounding boxes, the food items were also labeled with one or more of the NOVA groups depending on the level of processing the food items underwent. Since such task requires knowledge about the ingredients of food items and how they are processed, the NOVA dataset was fully annotated by expert Tunisian nutritionists. Our NOVA classification model described in this manuscript can thus assess the healthiness of Tunisian food items based on their images. However, our approach itself is general enough that it can be used to assess the healthiness of any food items, provided that accurately-labeled training data is obtained. Such data can be obtained using crowdsourcing as we did in the case of Tunisian food, as long as the annotators have sufficient knowledge about the food items in the images. Our NOVA classification model achieved an average F1-score of 0.86 on test data from the NOVA dataset.

Our main contributions in the manuscript can thus be summarized as follows:

1) we build a general deep-learning-based food item detection model that can be used to detect and localize food items in any food image,
2) we build a deep-learning-based multi-label classification model that can be used to classify a food item based on its image into one or more of the NOVA food groups, and
3) we provide a prototype to acquire training data for these two models using crowdsourcing.

The manuscript is organized as follows. In Section II, we give an overview of related work that addresses the problem of assessing the healthiness of food items using their images. In Section III, we describe our proposed deep learning approach to assess the healthiness of food items based on their images. Section IV describes the experiments we conducted to evaluate our proposed approach, their results and the error analysis of the proposed approach. Finally, we conclude and provide future directions in Section V.

II. RELATED WORK

There is a wealth of work on food analysis from food images. These works can be broadly categorized into works that focus on food item detection, works that focus on food healthiness assessment, or both.

A. FOOD ITEM DETECTION

Most of the works that aim to analyze food based on their images require algorithms and models for food item detection, recognition, and segmentation. For example, Akhi et al. [8] proposed a Convolutional Neural Network (CNN) model based on the ResNet-5 pre-trained model [9], which was used to extract features from fast-food images. The extracted features were then used to train a multi-class Support Vector Machines (SVM) classifier that classifies food images into 10 classes. Similalry, Liu et al. [10] proposed a deep learning approach based on CNNs that classifies food images that are captured in the real world. Aguilar et al. [11] developed a framework that addresses the problem of automatic food-tray analysis in restaurants. Their framework is based on CNNs, and is composed of food localization, recognition, and segmentation models. The first part of their framework is a food segmentation model that is based on a Fully Convolutional Network (FCN), and it aims to separate food items from the background (i.e, the tray). The second part of the framework then detects food items by using the YOLOv2 model [12].

Bolanos et al. [4] proposed another approach for generic simultaneous food localization and recognition. First, they trained the GoogleNet CNN model [13] to distinguish between food and non-food images. Second, they enhanced the previous model by adding Global Average Pooling (GAP) layer that aims to generate heat maps of food probabilities. Finally, bounding boxes were generated for the regions with a probability above a certain threshold. After detecting the food items, they fine-tuned the GoogleNet model to classify the items into various types. In addition to that, they built the EgocentricFood dataset, which contains food images that were captured using wearable cameras.

Unlike most of the approaches described above, *our food item detection model* proposed as part of the approach described in this manuscript does not make any assumptions about how the food images were captured or what they contain. It is thus able to detect food items of different shapes, sizes, and types in images that are taken in real settings, and with various resolutions and qualities. Our model is based on a customized version of YOLOv3 [3] that is able to detect food items on three different image scales with very high accuracy as indicated by our experiments.

B. FOOD HEALTHINESS ASSESSMENT

Assessing the healthiness of food items present in an image is a challenging computer vision task. Most works that address such problem rely on estimating the amount of calories in the food items to assess their healthiness. For instance, Liang et al. [14] proposed a calorie estimation approach that takes two images as an input: a top and a side view of a food item that include on its side a coin, which is used as a calibration object. They used a Faster r-CNN model [15] to detect food items using bounding boxes. They then applied image segmentation on the detected food items

for background removal using the GrabCut algorithm [16]. The segmented images are then used to estimate the volume and mass of the detected food items, which are in turn used to estimate the amount of calories in each food item.

Similarly, Myers et al. [17] developed the Im2Calories system that estimates the amount of calories in food dishes. They started by training a GoogLeNet model on the Food101 multi-labeled dataset [18]. They then used the DeepLab system [19] for semantic image segmentation to localize food items and segment them. Using the voxel representation and the segmentation mask of food items, they estimated the volume of each food item, and consequently predicted the amount of calories using the calorific density of each type of food. The authors, however, faced the problem of lack of sufficient calorie-annotated training data and thus could not do extensive evaluation of their approach because the texture properties and the color of the images in the Food101 dataset are different from the ones of real food images.

Another related work is the one by Lu et al. [20], where the authors proposed an AI system that is able to estimate the nutrient intake of hospital patients. They built a dataset consisting of 660 images by setting up a table that contains a camera on the top with a specific distance from the food items. In addition to that, they created a database that contains the recipes and the nutrient intake of the consumed meals. They used a Multi-Task Fully Convolutional Network (MTFCN) model [21] for image segmentation that aims to estimate the volumes of food items, which in turn helped in estimating the nutrient intakes based on the created database.

Gao et al. [22] proposed MUSEFood, which is a food-volume estimation approach that is different from all of the previous volume-estimation approaches. Their proposed approach does not require any training using food images with their corresponding volume information, and in addition eliminates the need to place a reference object of known size when capturing the images. Instead, they used microphones and speakers to calculate the vertical distance from the camera to the food items, which helped in estimating the actual volume of the food items and in turn estimating the amount of calories in the food items.

Chokr and Elbassuoni [23] also proposed an approach for calories estimation from food images. Their approach uses a machine learning model to predict the type of a food item in an image based on the image visual features. Their approach also predicts the size of the food item (in grams) and then based on these two predicted values as well as the original features of the image, it estimates the amount of calories in the food item. However, the authors only trained and tested their model on images that contain a single food item that belongs to only one of six different categories (burger, chicken, doughnut, pizza, salad and sandwich).

Overall, using volume and calories estimation approaches for assessing the healthiness of food items has many limitations including 1) the fact that the images of the food items should be captured from specific angles, 2) the need for reference objects, which are used in volume estimation, to be present in the food images, 3) training these models typically requires a large number of annotated images for each food type, and 4) there should be a specific predefined database that contains the nutrient information of the food items that exist in the images.

Sudo et al. [24] proposed a different healthiness assessment approach that is based on a feature extraction deep learning model that is followed by a ranking algorithm. First, they built a dataset of 850 images of meals that were taken from a top view. These images were ranked by registered nutritionists based on the healthiness of the whole meal from best to worst. Second, they built a feature extraction model that uses a CNN followed by a pyramid scene parsing network (PSPNET) [25], which outputs pixel-based feature maps. The extracted features were then used as an input to the ranking algorithm that uses another CNN. However, the authors reported that the correlation coefficient between the rankings of the nutritionists and the ground truth rank that is based on the nutritional facts of the meals was relatively low. The authors explained that their approach did not perform well because assessing the healthiness of food items by ranking them from best to worst without a specific criteria is not highly correlated with the ground truth healthiness of the food items.

Our healthiness assessment approach we propose in this manuscript addresses all the limitations of the above described approaches by utilizing a qualitative approach rather than a quantitative one. Instead of estimating the amount of calories in food items depicted in images, it classifies the food items into one or more of the four NOVA food groups based on their processing level. Our NOVA classification model can thus be used to accurately assess the healthiness of multiple food items in generic images without making any unrealistic assumptions about how the food images were taken. Moreover, our model does not require extensive annotation efforts as is the case with most of the above surveyed approaches.

III. APPROACH

In this section, we describe our end-to-end deep learning approach that can assess the healthiness of multiple food items in images. Our approach consists of two models: 1) a food item detection model that detects and localizes food items in an image, and 2) a classification model that classifies a detected food item into one or more of the four NOVA food groups. We describe each model separately next.

A. FOOD ITEM DETECTION MODEL

Our food item detection model is a customized object detection model that localizes food items in an image using bounding boxes. The food items can be of any type, shape, and size. Moreover, they can be in a dish, bowl, cup, or held by a person, etc. We first describe the data we used to train such a model, then we describe the architecture of the model itself afterwards.

FIGURE 2. Training example for the food item detection model.

TABLE 1. EgocentricFood dataset: distribution of food items across categories.

Category	Number of Items
Glass	975
Cup	775
Jar	37
Can	176
Mug	1,063
Bottle	2,250
Dish	983
Basket	96
Other	939

1) DATASETS

Training our food item detection model requires a dataset consisting of images that contain food items localized using bounding boxes. For example, Figure 2 shows a sample training example consisting of an image with multiple food items and where the food items are localized using bounding boxes. We used three different datasets to train our food item detection model. The first two are the EgocentricFood dataset [4] and the UECFood-256 dataset [5], which are both publicly available datasets, and the third is a custom dataset that we created ourselves and that contains food images taken using wearable cameras and in which food items were localized via crowdsourcing. The reason we used three different datasets is to ensure 1) we use a sufficiently large amount of data to train a deep-learning model, which usually require large amounts of data, and 2) we have sufficient variability in the training data so that our model is able to detect and localize all food items in any image regardless of their shape, type or size and regardless of how or where the image was taken.

Our first dataset, the EgocentricFood dataset, includes 5,038 images that were taken by wearable cameras. The images contain 7,294 different food items that are localized using bounding boxes. The dataset contains nine categories of food items, which are glasses, cups, jars, cans, mugs, bottles, dishes, baskets and others (food items that do not belong to any of the other categories). The distribution of food items across these categories is shown in Table 1. Figure 3 shows a sample of images from this dataset.

FIGURE 3. Sample images from the EgocentricFood dataset.

TABLE 2. UECFood-256 dataset: Top-10 categories with the highest number of items.

Category	Number of Items
Miso Soup	728
Rice	620
Ramen Noodle	353
Green Salad	342
Beef	246
Hamburger	233
Egg	224
Toast	218
Fried Rice	169
French Fries	153

FIGURE 4. Sample images from the UECFood-256 dataset.

Our second dataset, the UECFOOD-256 dataset, is an Asian food dataset that consists of food images crawled from the Web. Similar to the first dataset, the food items in this dataset are also localized in the images using bounding boxes. The dataset consists of a total of 28,898 images that contain 31,395 food items belonging to 256 different categories. The dataset contains at least 100 images for each category. Table 2 shows the 10 categories with the highest number of items in the dataset. Figure 4 shows sample images from the UECFOOD-256 dataset. Since the dataset consists of images crawled from the Web, some of these images are not taken in real-life settings such as commercial images of food items.

FIGURE 5. Sample images from the NOVA dataset.

The third and final dataset we used to train our food item detection model, which we refer to as the NOVA dataset, consists of 1,800 food images. The images contain various Tunisian food items as they were taken by school children in Tunisia using wearable cameras. Figure 5 shows sample images from the NOVA dataset. To localize the food items in the images, we created a custom interface on the crowdsourcing platform Labelbox[3] that allows annotators to localize each food item in a given image by drawing a bounding box around it. Each image was annotated by two different trained nutritionists. After all the images in the NOVA dataset were annotated on Labelbox, we ended up with 4,201 localized food items in the 1,800 images.

2) MODEL

Our generic food item detection model is based on YOLOv3 model that was developed by Redmon and Farhadi [3]. YOLOv3 is a one stage real-time object detection model that localizes general objects in images and videos using bounding boxes. This model has been shown to outperform other object detection models in terms of both detection quality and time [3].

Similar to YOLOv3, our food detection model is made up of two main components, which are a feature extractor and a feature detector. The feature extractor is a CNN referred to in YOLOv3 as Darknet-53. It is made up of 53 layers (hence the name Darknet-53) with 3×3 and 1×1 convolutional layers followed by residual connections [9]. 53 additional layers are also added to the Darknet-53 network that serve as a detection head, resulting in a total of 106 convolutional layers. The detection head of the model performs object detection on three different image scales by applying 1×1 detection kernels on their corresponding feature maps [3]. The three scales of each image are determined by the stride parameters in the CNN, which are responsible for down-sampling the

[3]https://labelbox.com/

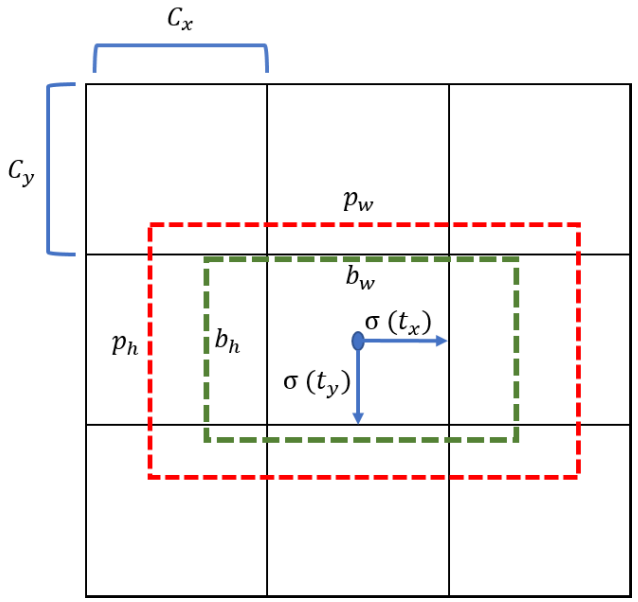

FIGURE 6. Bounding box coordinates prediction.

images by factors of 32, 16, and 8, respectively. Since all the images in our three training datasets have a resolution of 416×416, we ended up with three different resolutions for each image: 52×52, 26×26, and 13×13. This technique of performing object detection on three different scales of each image helps in improving the accuracy of detecting food items of different sizes as we show in our experiments in Section IV.

After getting the feature maps, the input image is divided into $S \times S$ grid according to the extracted feature map size. For example, a 416×416 image with a 26×26 feature map will result in an image divided into 26×26 cells. Each of the cells predicts three bounding boxes, objectness scores (i.e., the probability that there is an object in a bounding box), and the classes the detected objects belong to. The model outputs a bounding box coordinate (t_x, t_y, t_w, t_h), where (t_x, t_y) is the center of the bounding box and (t_w, t_h) is the width and height of the box. The bounding boxes are calculated with the help of the anchor boxes, which are predefined bounding boxes that are used to predict the bounding boxes coordinates by predicting the offsets to the anchor boxes. Figure 6 shows the predicted bounding box coordinates in green and the anchor box in red.

The anchor boxes are calculated using the k-means clustering algorithm [26], which starts by choosing k random points as initial clustering centroids. It then calculates the distance from each point to each of the centroids, and finally assigns each point to its nearest centroid. The algorithm then proceeds to update the centroids until the algorithm converges (i.e., the centroids do not change anymore). In our model, the input to the clustering algorithm is the widths and heights of the bounding boxes and we set $k = 9$ since we need three anchor boxes per each of the three image scales. To calculate the distance from a centroid to a bounding box,

we subtract 1 from the Intersection over Union (IoU) of the box and the centroid as shown in Equation 1:

$$distance(Box, Centroid) = 1 - IoU(Box, Centroid) \quad (1)$$

where IoU is a measure that calculates the similarity between two bounding boxes using Jaccard index by dividing the intersection of the shapes by their union as shown in Eq. 2:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

To calculate the bounding boxes coordinates, the model first transforms the output of the CNN (t_x, t_y, t_h, t_w) to (b_x, b_y, b_w, b_h), where (b_x, b_y) is calculated by applying sigmoid function (Eq. 3 and Eq. 4) on the predicted (t_x, t_y) and adding (c_x, c_y), which is the top-left offset of our grid from the current cell of the feature map:

$$b_x = \sigma(t_x) + c_x \quad (3)$$

$$b_y = \sigma(t_y) + c_y \quad (4)$$

(b_w, b_h) is the width and height of the predicted bounding box, which are calculated using (p_w, p_h), which in turn is the anchor box's coordinates as can be seen in Eq. 5 and Eq. 6:

$$b_w = p_w e^{t_w} \quad (5)$$

$$b_h = p_h e^{t_h} \quad (6)$$

In addition to the bounding box coordinates, the model outputs an objectness score, which is calculated using logistic regression (Eq. 7) and indicates the probability that there is an object inside a certain bounding box:

$$\sigma(x) = \frac{1}{(1 + e^{-x})} \quad (7)$$

Moreover, the model also predicts classes for the detected objects using a sigmoid function (i.e., multi-label classification where the model can predict more than one class per bounding box). Note that in our case, all objects are assumed to belong to one class, namely food, as compared to the general YOLOv3 model, which is typically used to detect objects that belong to multiple classes (e.g., car, pedestrian, truck, tree, etc.). That is, in our case, the objectness score represents the probability that there is a food item inside a bounding box, and the classification is a binary one (i.e., a detected object is either a food item or not).

The YOLOv3 model calculates the bounding box error using Mean Squared Error (MSE) of $t - \hat{t}$ [3], where t is the ground truth coordinates, and $\hat{t}$ is the predicted ones. In our model, we use the Generalized IoU (GIoU) proposed by Rezatofighi et al. [27] as a loss function. GIoU is an extension of IoU that addresses its limitations as pointed out in [27]. First, If $|A \cap B| = 0$, then $IoU = 0$ and therefore IoU will not reflect if the bounding boxes are near or far from each other. Second, IoU does not actually reflect the overlap between the bounding boxes. The GIoU metric was proposed to solve these problems and the evaluation in [27] shows that using the GIoU loss improves the performance of many object detection models, including YOLOv3, on popular object detection benchmarks such as the COCO dataset [28]. GIoU is calculated using Eq. 8:

$$GIoU(A, B) = IoU(A, B) - \frac{|C \backslash (A \cup B)|}{|C|} \quad (8)$$

where C is the smallest box enclosing A and B, and $|C \backslash (A \cup B)|$ calculates the area occupied by C without A and B.

The values of IoU are in the range $[0, 1]$, whereas the values of GIoU are in the range $[-1, 1]$, where 1 is the maximum value when two bounding boxes overlap and -1 is the minimum value when the bounding boxes are not overlapping. GIoU loss is calculated by subtracting 1 from the value of GIoU as shown in Eq. 9:

$$\mathcal{L}_{GIoU} = 1 - GIoU \quad (9)$$

Unlike the standard YOLOv3, which uses BCE loss for objectness scores and class prediction, we use BCE with Logits Loss (BCEWithLogitsLoss) as shown in Eq. 10:

$$\begin{aligned} \text{BCEWithLogitsLoss} = -\frac{1}{n} \times \sum_i & \left(y_i \times \log\left(\sigma\left(\hat{y}_i\right)\right)\right. \\ & \left. + (1 - y_i) \times \log\left(1 - \sigma\left(\hat{y}_i\right)\right)\right) \end{aligned} \quad (10)$$

where y is the true label of the image, $\hat{y}$ is the predicted probability, and σ is the sigmoid function that maps the values between 0 and 1. This is a more stable version of BCE loss, which takes too long to converge compared to BCEWithLogitsLoss that uses sigmoid before applying the BCE loss, resulting in more stable results [29].

Our final loss function that we need to minimize while training the model is thus a sum of the object (i.e., food item) localization loss, the classification loss (whether a localized object is a food item or not) and the objectness loss (the probability of a food item being present in a bounding box) as shown in Eq. 11:

$$\mathcal{L}_{model} = \mathcal{L}_{Localization} + \mathcal{L}_{Classification} + \mathcal{L}_{Objectness} \quad (11)$$

where:

$$L_{Localization} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathcal{L}_{GIoU}(b_i^j, \hat{b}_i^j) \quad (12)$$

$$\begin{aligned} L_{Objectness} = & \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} \left[c_i^j \times \log\left(\sigma\left(\hat{c}_i^j\right)\right)\right. \\ & \left. - \left(1 - c_i^j\right) \times \log\left(1 - \sigma\left(\hat{c}_i^j\right)\right)\right] \\ & + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{noobj} \left[c_i^j \times \log\left(\sigma\left(\hat{c}_i^j\right)\right)\right. \\ & \left. - \left(1 - c_i^j\right) \times \log\left(1 - \sigma\left(\hat{c}_i^j\right)\right)\right] \end{aligned} \quad (13)$$

TABLE 3. NOVA dataset: distribution of food items over the NOVA food groups.

NOVA Groups	1	2	3	4	1-2	1-3	1-4	2-3	3-4	1-2-3	1-2-4	1-3-4
Number of Food Items	1,094	122	759	557	510	40	4	7	11	569	49	6

$$L_{Classification} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{obj} \sum_{c \in class} [p(c_i^j) \times \log\left(\sigma\left(p(\hat{c}_i^j)\right)\right) - \left(1 - p(c_i^j)\right) \times \log\left(1 - \sigma\left(p(\hat{c}_i^j)\right)\right)] \quad (14)$$

and $s^2 = (s \times s)$ is the number of cells of the feature map, and B, which is set to 3 in our model, is the number of bounding boxes generated by each cell. In the localization loss equation (Eq. 13), b_i^j and $\hat{b}_i^j$ are the true and predicted bounding boxes coordinates, respectively. The objectness loss (Eq. 13) is calculated using BCEWithLogitsLoss, where c_i^j and $\hat{c}_i^j$ are the true and predicted confidences, respectively. $\mathbb{1}_{i,j}^{noobj}$ is used to determine if the j^{th} bounding box of the i^{th} cell is not responsible for the detection of the object. In addition, λ_{noobj} is the weight of GIoU loss, which is set to 0.5 in our experiments. Similarly, the classification loss (Eq. 14) is calculated using BCEWithLogitsLoss, where $p(c_i^j)$ is the ground truth probability that the object in the i^{th} cell belongs to class c (in our case the single class food), and $p(\hat{c}_i^j)$ is the predicted probability. Similar to the objectness loss, $\mathbb{1}_{i,j}^{obj}$ checks if the j^{th} bounding box of the i^{th} cell is the one responsible for the detection. It is equal to 1 if the $GIoU(BoundingBox, GroundTruth)$ is 1 (i.e., the maximum value possible), and 0 otherwise.

B. NOVA CLASSIFICATION MODEL

Once food items have been detected and localized in an image using our food item detection model described above, they are fed to the NOVA classification model, which classifies each food item into one or more of the four NOVA groups [1] based on their nature and the extent of the industrial processes that were applied to the food items. The first group is the Unprocessed or Minimally Processed Food, which includes natural food items such as vegetables, fruits, eggs, milk, water, etc. The second group is the Processed Culinary Ingredients, which is usually acquired from the first group and includes butter, oil, honey, etc. The third group is the Processed Food, which includes products made by adding salt, sugar or other Group 2 substances to Group 1 food such as unpackaged bread, canned fish, canned vegetables, etc. Finally, the fourth group is the Ultra-Processed Food, which includes food items that are produced using a series of industrial processes such as chips, chocolate, soft drinks, hotdog, etc. Since a food item can belong to more than one of these groups at the same time, our NOVA classification model is a multi-label classifier. We first describe the dataset we used to train the model and then describe the model itself afterwards.

1) DATASET

To train our NOVA classification model, we used the NOVA dataset we created ourselves as part of the training data for the food item detection model. To the best of our knowledge, no public datasets are available that contain images of food items with their corresponding NOVA food groups. Recall that the NOVA dataset consists of 1,800 images of Tunisian food items captured through wearable cameras in real-life settings. Each one of these images was annotated by two different trained Tunisian nutritionists using the Labelbox crowdsourcing platform. The annotators localized each food item in each image by drawing a bounding box around it. In addition, for each localized food item, each of the two annotators assigned it to one or more of the four NOVA food groups based on its processing level as perceived by the annotator. In case it was not possible for an annotator to assign a certain food item to any of the NOVA groups based on its image, the annotator indicated this by not assigning the food item to any of the groups.

The agreement between the two annotators was then calculated over all the food items detected in all the images in the NOVA dataset using 1) the IoU of the bounding boxes provided by the two annotators to localize each detected food item, and 2) the agreement over the assigned NOVA groups between the two annotators for each detected food item. The agreement score between the two annotators was 85%. As can be observed, there was a disagreement on the food groups for a small portion of the detected food items between the two annotators, which can be mainly attributed to the fact that some food items contain some ingredients that were not visible in the images such as oil and salt. To address this, the two annotators were asked to go over all the food items with an agreement less than 95% until they agreed on their NOVA groups.

After the just described annotation process was completed, we extracted cropped images of all the detected food items in the original images using their bounding boxes' coordinates. We thus ended up with a dataset that contains 4,201 images of different food items that belong to the different NOVA groups. Out of these, 185 food items were not assigned to any NOVA food group, and were thus removed from the dataset. We also removed any images of food items that were too blurry or too small from the dataset. Thus, the final dataset that we used to train the NOVA classification model consisted of 3,728 images of food items that belong to the different NOVA groups as shown in Table 3. As a food item can belong to multiple NOVA groups at the same time, in the table, the column corresponding to xy indicates the number of food items that belong to both groups x and y. For example, the number of food items in the NOVA

dataset that belong to groups 1 and 2 is equal to 510 food items.

2) MODEL

Our NOVA classification model is based on the MobileNetV2 architecture [7]. The architecture of MobileNetV2 contains an initial fully convolutional layer with 32 filters, followed by 19 residual bottleneck layers. It uses ReLU6 as a non-linearity because of its robustness when used with low-precision computations. It uses a kernel size of 3×3 as is standard in modern networks, and utilizes dropout and batch normalization during training. To adapt MobileNetv2 to our case and use it to classify a food item into one or more of the NOVA groups, we built a model that is made up of MobileNetV2 architecture loaded with frozen ImageNet weights, and we added a classifier on the top of it. The classifier consisted of a global average pooling layer followed by a dense layer of 250 neurons and a dropout layer with a dropout rate of 0.5. The output layer is a dense layer with four neurons representing the four NOVA groups. Each of the four neurons uses a sigmoid activation function to output the probability of a food item belonging to each of the four NOVA classes. Finally, to train the full model, we used Binary Cross Entropy loss as a loss function.

IV. EXPERIMENTS

In this section, we describe how we trained our two models, the food item detection model and the NOVA classification model. We report on the performance of each model, and then provide some error analysis for each one.

A. FOOD ITEM DETECTION MODEL

For the food item detection model, we trained three different versions of the model described in Section III using our three training datasets, the UECFood256 dataset, the EgocentricFood dataset, and the NOVA dataset. For the first model, we obtained the pretrained weights for the base YOLOv3 model on the COCO dataset [28] and retrained the whole food item detection model using the UECFood256 dataset. The dataset was split into 80% for training (23,119 images) and 20% for validation (5,780 images). We refer to this model as the *UECFood256 model*.

For our second model, we started with the UECFood256 model, and froze the weights of the backbone network, Darknet-53, which is used as the feature extractor. We then only trained the head of our food item detection model via transfer learning using the EgocentricFood dataset. Again, this dataset was split into 80% for training (4,038 images) and 20% for validation (1,000 images). We refer to this second model as the *EgocentricFood model*.

Finally, for our third model, we again started with the UECFood256 model and did transfer learning by freezing the backbone of our model, Darknet-53. We then trained only the head of the model, however on a combination of the NOVA dataset and the EgocentricFood dataset, but after removing a random sample of 500 images from the NOVA dataset, which is used as our test data. We split the combined data into 80% for training and 20% for testing. We refer to this model as the *NOVA and EgocentricFood model*.

The three above-described models were all trained for 100 epochs on images of size 416×416. We used a learning rate of 0.01 and a decay weight of 0.0005. Moreover, we used different anchor boxes, depending on the dataset that was used for training. That is, we generated nine anchor boxes for each dataset using the k-means algorithm as explained in Section III. Table 4 shows the anchor boxes for each dataset on three scales.

To evaluate the three models, we used the mean Average Precision (mAP) metric. mAP is the mean of AP over the classes and since we only have one class, namely food, the mAP will be the same as AP. To compute mAP, we need to compute the precision (Eq. 15) and the recall (Eq. 16). We also need to specify an IoU threshold value. For a threshold value of 0.5, a detected object is a True Positive (TP) if $IoU \geq 0.5$, otherwise it is a False Positive (FP). On the other hand, a False Negative (FN) is a food item that was not detected and a True Negative (TN) is any part of the image that does not contain any food item.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

In our evaluation, we used two mAP versions: the Pascal Visual Object Classes (VOC) metric [30] and the COCO metric [28]. The Pascal VOC metric, $mAp@0.5$ calculates the mAP for $IoU \geq 0.5$. The COCO metric, $mAP@[0.5 : 0.95]$, calculates the average of the mAP over different IoU thresholds that range from 0.5 to 0.95 with a step size of 0.05.

In order to compare the three versions of our food item detection model describe above, we tested them on 500 images that were sampled from the NOVA dataset and were not part of the training or validation. Table 5 shows the results for the three models using different metrics. We can clearly see that the *NOVA and EgocentricFood model* outperforms the first two models for all metrics. The UECFood256 model achieved the worst results since the images in this dataset were crawled from the Web and are all homogeneous, which thus does not generalize well to other more realistic images such as those captured through wearable cameras. The EgocentricFood model achieved better results since the images in this dataset are taken from wearable camera, which are very similar to the test images.

As a baseline, we also trained a standard YOLOv3 model on the combined NOVA and EgocentricFood datasets (with the same train-validation split as the NOVA and EgocentricFood model) and compared it with our best performing model, the NOVA and EgocentricFood model. Table 6 shows the localization loss and the objectness loss for each model on the validation set. As can be seen from the table, the NOVA and EgocentricFood model outperformed the baseline YOLOv3 model for both loss functions (smaller

TABLE 4. The generated anchor boxes for each dataset.

Dataset/ Scale	Small (32)	Medium (16)	Large (8)
UECFood256	$(195 \times 174), (319 \times 259), (417 \times 369)$	$(535 \times 289), (507 \times 420), (581 \times 381)$	$(445 \times 549), (606 \times 454), (593 \times 574)$
EgocentricFood	$(68 \times 73), (82 \times 124), (182 \times 98)$	$(124 \times 149), (103 \times 233), (225 \times 179)$	$(173 \times 332), (426 \times 193), (402 \times 367)$
EgocentricFood+Nova	$(63 \times 54), (76 \times 104), (158 \times 87)$	$(98 \times 152), (107 \times 247), (177 \times 172)$	$(339 \times 151), (180 \times 320), (411 \times 303)$

TABLE 5. Test results for three food item detection models.

Model	Precision	Recall	mAP@[0.5:0.95]	mAP@0.5
UECFood256	0.44	0.22	0.11	0.22
EgocentricFood	0.70	0.73	0.41	0.71
NOVA and EgocentricFood	**0.87**	**0.91**	**0.65**	**0.90**

TABLE 6. Validation results for the YOLOv3 vs. the NOVA and EgocentricFood models.

Model	Localization Loss	Objectness Loss	mAP@0.5
YOLOv3	0.12	0.08	0.85
NOVA and EgocentricFood	**0.01**	**0.02**	**0.88**

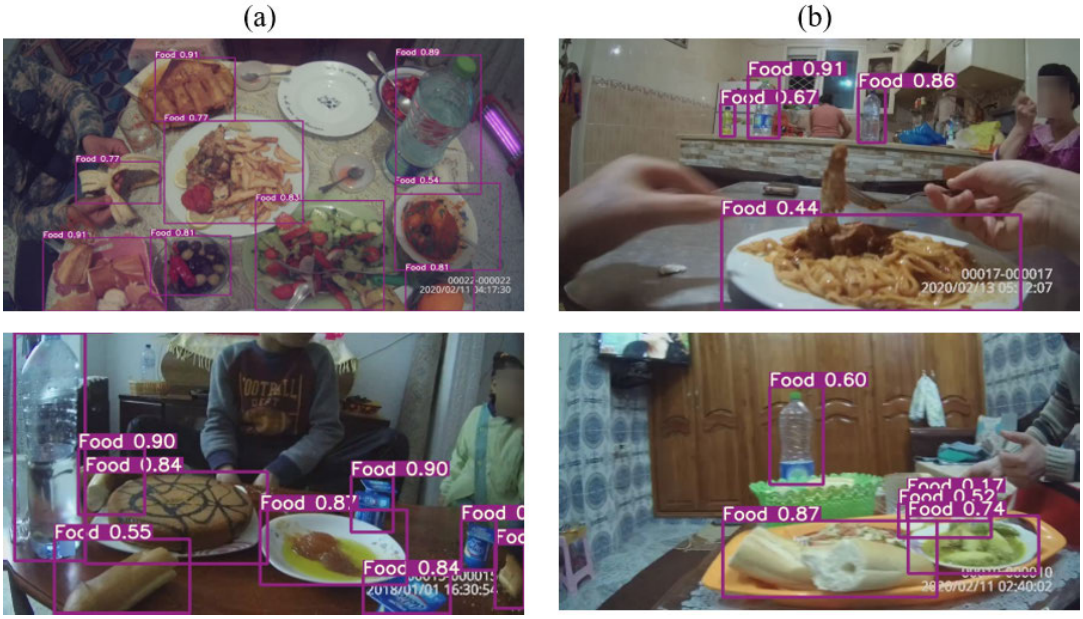

FIGURE 7. Results of the food item detection model on sample test images.

is better) as well as in terms of $mAP@o.5$ (with 3% improvement). This can be mainly attributed to the use of $\mathcal{L}_{GIoU}$ instead of MSE when calculating the error between the ground truth bounding boxes and the predicted ones.

Finally, in Figure 7, we show the results obtained by applying the NOVA and EgocentricFood model on a sample test images from the NOVA dataset. As can be seen from the figure, the model was able to detect most of the food items that appear in the sample images shown. On the other hand, the model was not able to detect some of the food items that are occluded by other objects such as the dish in the top image of set (a), which is occluded by a water bottle. On the other hand, overall the model was robust, as it was able to detect small food items that are even far from the table, as can be seen in the top image of set (b). This is mainly due to the fact that our food item detection model is applied on three different image scales (small, medium, and large). This actually had a negative effect on the reported precision, by increasing the number of false positives, since the annotators did not localize food items that are far away from the table when generating the ground truth. Finally, some of the food items had overlapping bounding boxes(such

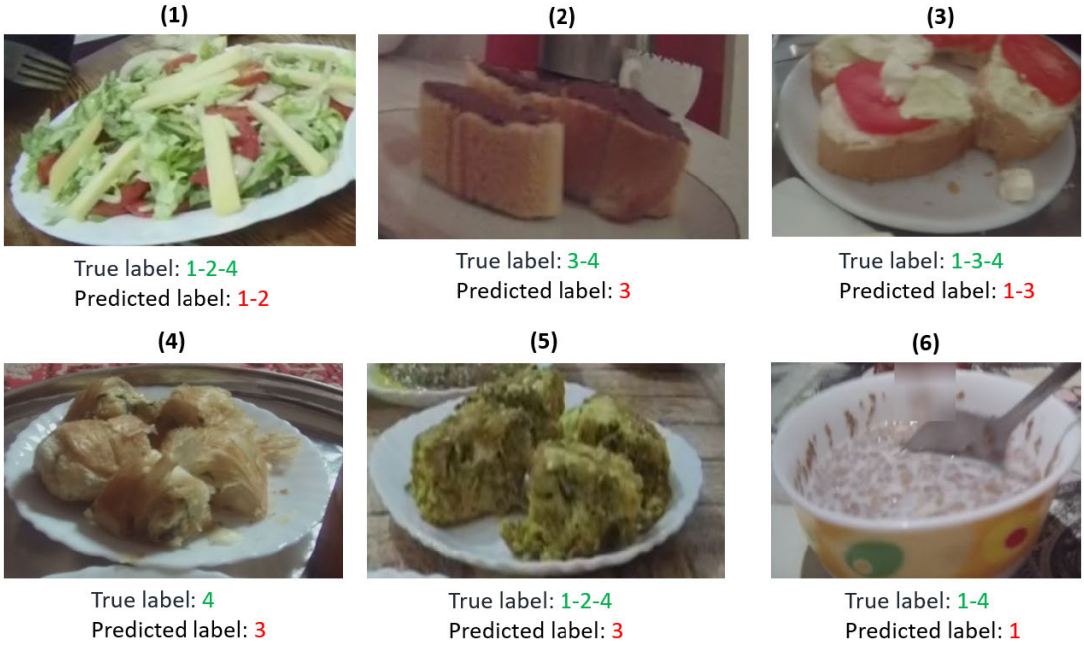

FIGURE 8. Sample misclassified food items by the NOVA classification model.

as the bottom image in set (b)), and we addressed this by excluding the ones with a low confidence score (less than 0.40), as is custom in object detection in general.

B. NOVA CLASSIFICATION MODEL

The NOVA classification model was trained using the NOVA dataset described in Section III. The dataset was split into 80% for training, 10% for validation and 10% for testing. Recall that the model consists of two parts, a base MobileNetV2 backbone and a classifier on top of it that consisted of a global average pooling layer followed by a dense layer of 250 neurons and a dropout layer with a dropout rate of 0.5. The final output layer of the model is a dense layer with four neurons representing the four NOVA groups. To train the model, the MobileNetV2 backbone was loaded with pretrained ImageNet weights (i.e., transfer learning), its layers were frozen and the rest of the model was trained for 20 epochs using the Adam optimizer with a learning rate of 0.001. After that, we did fine-tuning by unfreezing the last 55 layers of the model and retraining the model for 10 more epochs with a learning rate of 0.0001.

Since the NOVA dataset contains images of various sizes, we used different image sizes as a hyperparameter (128, 160, 192, 224). We resized the images using Bilinear Interpolation, which is a resampling method that calculates a new pixel value based on the distance weighted average of the nearest four pixels [31]. After training our NOVA classification model with different image sizes, the 224×224 image size was the best fit for the model based on the validation set. The results of this model on the testing data is shown in Table 7, which shows the precision, the recall, and the F1-score for each of the four NOVA groups.

TABLE 7. NOVA classification model test results.

Group	Precision	Recall	F1-score
1	0.92	0.86	0.89
2	0.90	0.85	0.87
3	0.86	0.84	0.85
4	0.84	0.85	0.84
Average	**0.88**	**0.85**	**0.86**

While the overall performance of the model on the test data was relatively high, however, some of the images were misclassified due to the complexity of the food items. We noticed that our model was not able to predict all the ground truth NOVA groups for some of the images that contain ingredients that are not visible to the model such as salt and oil. Figure 8 shows a sample of misclassified images by the NOVA model. For example, the first image's ground truth NOVA groups are 1 and 2 since it is a salad, and 4 because it contains cheese. The model was able to correctly predict that the food item belong to groups 1 and 2, however, it did not predict group 4 since most of the salad related images in the training set belong to groups 1 and 2 only. Another example is image 3, where the ground truth indicates that the food item belongs to the NOVA groups 1, 3, and 4 since it contains bread, tomatoes, and cheese. The model correctly predicted that it belongs to groups 1 and 3 and it missed group 4. These results explain why we did not obtain a very high recall for some of the NOVA groups.

V. CONCLUSION AND FUTURE WORK

In this manuscript, we proposed DeepNOVA, a novel end-to-end deep learning approach that can assess the healthiness

of food in images based on the NOVA classification system. Our approach can be used by nutritionists to study the dietary intake of a target population, which is traditionally done using interviews and questionnaires, and is known to suffer from recall bias. Our approach consists of two main models, a food item detection model followed by a NOVA classification model. The food item detection model is a custom object detection model that is trained to detect and localize any food item in an image, while the NOVA classification model is a multi-label classification model that assigns a food item to one or more of the NOVA food groups based on its processing level. Both models were trained using a combination of public datasets and a custom dataset that we generated using wearable cameras and that was annotated by trained nutritionists.

In future work, we plan to use DeepNOVA in real-world nutritional case studies. Since the NOVA classification model was trained on images of Tunisian food, and was labeled by Tunisian nutritionist, the model has to be fine-tuned based on the type of food in the case study. We do not consider this a limitation as it mimics human behavior when assessing the healthiness of food, which requires local knowledge about the food being assessed and how it is processed. However, our model can be used as a pretrained model and fine-tuned using other food types via transfer learning.

ACKNOWLEDGMENT

The authors would like to thank the School and Community Drivers of Child Diets in Arab Cities; Identifying Levers for Intervention SCALE Research Group for supporting this work (Akik Chaza, Doggui Radhouene, El-Helou Nehmat, Jamaluddine Zeina, Safadi Gloria, and Trabelsi Tarek).

REFERENCES

[1] C. A. Monteiro, G. Cannon, M. Lawrence, M. D. C. Louzada, and P. P. Machado, "Ultra-processed foods, diet quality, and health using the NOVA classification system," FAO, Rome, Italy, vol. 48, 2019. [Online]. Available: https://www.fao.org/3/ca5644en/ca5644en.pdf

[2] M. M. Lane, J. A. Davis, S. Beattie, C. Gómez-Donoso, A. Loughman, A. O'Neil, F. Jacka, M. Berk, R. Page, W. Marx, and T. Rocks, "Ultraprocessed food and chronic noncommunicable diseases: A systematic review and meta-analysis of 43 observational studies," *Obesity Rev.*, vol. 22, no. 3, Mar. 2021, Art. no. e13146.

[3] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[4] M. Bolanos and P. Radeva, "Simultaneous food localization and recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3140–3145.

[5] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 3–17.

[6] C. A. Monteiro and A. Astrup, "Does the concept of 'ultra-processed foods' help inform dietary guidelines, beyond conventional classification systems? YES," *Amer. J. Clin. Nutrition*, vol. 2022, pp. 1–45, Jun. 2022.

[7] 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.

[8] A. B. Akhi, F. Akter, T. Khatun, and M. S. Uddin, "Recognition and classification of fast food images," *Global J. Comput. Sci. Technol.*, vol. 18, pp. 1–8, Jul. 2018.

[9] 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.

[10] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment," in *Proc. Int. Conf. Smart Homes Health Telematics*. Cham, Switzerland: Springer, 2016, pp. 37–48.

[11] E. Aguilar, B. Remeseiro, M. Bolaños, and P. Radeva, "Grab, pay, and eat: Semantic food detection for smart restaurants," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3266–3275, Dec. 2018.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[13] 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.

[14] Y. Liang and J. Li, "Deep learning-based food calorie estimation method in dietary assessment," 2017, *arXiv:1706.04062*.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[16] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut' interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[17] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1233–1241.

[18] 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.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[20] Y. Lu, T. Stathopoulou, M. F. Vasiloglou, S. Christodoulidis, Z. Stanga, and S. Mougiakakou, "An artificial intelligence-based system to assess nutrient intake for hospitalised patients," *IEEE Trans. Multimedia*, vol. 23, pp. 1136–1147, 2021.

[21] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[22] J. Gao, W. Tan, L. Ma, Y. Wang, and W. Tang, "MUSEFood: Multi-sensor-based food volume estimation on smartphones," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 899–906.

[23] M. Chokr and S. Elbassuoni, "Calories prediction from food images," in *Proc. 29th IAAI Conf.*, 2017, pp. 4664–4669.

[24] K. Sudo, K. Murasaki, T. Kinebuchi, S. Kimura, and K. Waki, "Machine learning–based screening of healthy meals from image analysis: System development and pilot study," *JMIR Formative Res.*, vol. 4, no. 10, 2020, Art. no. e18507.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[26] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Phys. Proc.*, vol. 25, pp. 1104–1109, Jan. 2012.

[27] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[29] PyTorch. (2019). *Bcewithlogitsloss*. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

[30] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[31] G. R. Arce, J. Bacca, and J. L. Paredes, "Nonlinear filtering for image analysis and enhancement," in *The Essential Guide to Image Processing*. Amsterdam, The Netherlands: Elsevier, 2009, pp. 263–291.

SHADY ELBASSUONI received the Ph.D. degree from the Max-Planck Institute for Informatics (MPII), Germany. He is currently an Associate Professor with the Computer Science Department, American University of Beirut. Before joining AUB, he was a Postdoctoral Researcher at the Qatar Computing Research Institute (QCRI). He has over 60 publications in these areas in top-tier conference proceedings and journals. His current research spans multiple areas including machine learning, information retrieval, and crowdsourcing. His research interests include searching and ranking for knowledge graphs and applications of machine learning particularly deep learning in various domains including natural language processing, computer vision, information retrieval, public health, and medicine. His research work also addresses problems related to fairness in algorithmic decision making and crowdsourcing.

HALA GHATTAS is currently a Public Health Nutritionist whose research focuses on the links between inequity, food insecurity, nutritional status, and health. Her work explores the social and structural determinants, and health consequences of both under and over-nutrition in the contexts of the global nutrition transition and regional conflicts in the middle east. She has developed novel tools to measure food environments and food insecurity experience in low and middle-income settings. She has also led the nutrition and health components of multidimensional poverty surveys and vulnerability assessments; and designed and evaluated public health programs to address the overlapping burdens of food insecurity, malnutrition, and chronic diseases particularly in refugee populations.

JALILA EL ATI is currently a Professor of nutrition, a Researcher on public health nutrition, the Head of the Studies and Planning Department, and responsible of the LR Nutritional Surveillance and Epidemiology in Tunisia—SURVEN. She has a research experience more than 30 years in three major domains: physiology, nutrition, and food sciences. She has led a number of large epidemiological surveys to assess nutritional status, food customs, food consumption, health problems, and their determinants. She was a member of the WHO Global Coordination Mechanism on the Prevention and Control of NCDs (WHO GCM/NCD), a member of the Working Group WHO Guideline Development Group—Nutrition Actions (WHO/HQ), and a member of the WHO Steering Committee to elaborate the WHO Guideline Management of the adolescents' obesity. She is the Principal Investigator of international projects funded by FAO, UNICEF, WHO, USAID, ANR, ENI CBC MED, GAIN, PAM, and IDRC.

ZOULFIKAR SHMAYSSANI received the M.S. degree in computer science from the American University of Beirut. He is currently a Research Assistant with the American University of Beirut. His research interests and expertise include machine learning and computer vision.

SARAH KATERJI received the M.S. degree in computer science from the American University of Beirut. She is currently a Research Assistant with the American University of Beirut. Her research interests and expertise include machine learning and computer vision.

YORGO ZOUGHBI received the M.S. degree in computer science from the American University of Beirut. He is currently a Research Assistant with the American University of Beirut. His research interests and expertise include machine learning and computer vision.

ALINE SEMAAN is currently pursuing the Ph.D. degree with the Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium. Her main research interests include reproductive and maternal health.

CHRISTELLE AKL is currently pursuing the Ph.D. degree in epidemiology with the American University of Beirut. She has a background in nutrition and public health. Her main research interests include dietary assessment, nutritional outcomes, and non-communicable diseases in the arab region.

HOUDA BEN GHARBIA is currently pursuing the Ph.D. degree in biology. She is currently a Researcher on public health nutrition. She is also a member of the Research Laboratory Nutritional Surveillance and Epidemiology in Tunisia—SURVEN. She has research experience in nutrition, epidemiology, and food science. She participated in several epidemiological surveys to assess nutritional status, food customs, food consumption, health problems, and their determinants.

SONIA SASSI received the master's degree in epidemiological and statistics from the Faculty of Medicine of Tunis and the Ph.D. degree in biology from the Faculty of Science of Tunis. She has experience working on several projects focusing on nutrition epidemiology and food science. She is currently a Researcher on public health nutrition and a member of the Research Laboratory Nutritional Surveillance and Epidemiology in Tunisia—SURVEN. She also works with the Studies and Planning Department, National Institute of Nutrition.