Hindawi
Journal of Food Processing and Preservation
Volume 2023, Article ID 6612302, 15 pages
https://doi.org/10.1155/2023/6612302

Research Article

Hybrid Deep Learning Algorithm-Based Food Recognition and Calorie Estimation

Ritu Agarwal,[1] **Tanupriya Choudhury**,[1,2] **Neelu J. Ahuja**,[1] **and Tanmay Sarkar**[3]

[1] *School of Computer Science, University of Petroleum and Energy Studies (UPES), Bidholi Campus, Dehradun, Uttrakhand 248007, India*
[2] *CSE Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India*
[3] *Department of Food Processing Technology, Malda Polytechnic, West Bengal State Council of Technical Education, Government of West Bengal, Malda 732102, West Bengal, India*

Correspondence should be addressed to Tanmay Sarkar; tanmay@wbscte.ac.in

Received 28 August 2023; Revised 17 October 2023; Accepted 18 October 2023; Published 24 November 2023

Academic Editor: Adadi Parise

Every individual requires some sort of system that informs them about portions and calories of food, as well as providing them with directions on how to consume it. In our study, we propose a hybrid architecture that makes use of deep learning algorithms to forecast the number of calories in various food items on a bowl. This consists of three major components: segmentation, classification, and calculating the volume and calories of food items. When we use a Mask RCNN, the images are first segmented. Using the YOLO V5 framework, features are collected from the segmented images and the food item is categorized. In order to determine the dimensions of each food item, we identify the items first. In order to calculate the quantity of the food item, the estimated dimension must be used. The calories are then computed using the food item's volume. The aforementioned approaches, which were trained on the dataset's food images, that correctly identified and forecasted a food item's calories had an accuracy of 97.12%. To Provide directions on how to consume food is expected by individual and will be completed after knowing intake of volume of food.

1. Introduction

A lot of individuals enjoy consuming fast food and soft drinks, which are high in calories and sugar. Obesity levels are rising as a result of people getting less exercise, not knowing enough about nutrition, and having uncontrollable eating habits. Obesity-related conditions include hypertension, diabetes, heart problems, and breathing difficulties. Due to excess body weight, obesity leads to persons having knee or other joint ligament injuries [1]. While walking or climbing stairs, they also experience breathing difficulties, and their hearts work too hard to pump blood throughout the body. A diabetic is someone who has a condition that causes their body to produce less insulin, which raises their blood sugar levels [2]. A calorie imbalance between caloric intake and caloric expenditure is what is known as obesity and overweight [3]. There have been a lot of recent incidences of obesity among young people. Young people, particularly kids, are more prone than ever to eat fast food and other fast-food items that raises their chance of contracting a number of serious ailments including cardiac blockage, stroke, diabetes, issues with the liver, and renal damage due to the impact of western way of life, the web, and TV commercials [4, 5].

Obesity is viewed as a global epidemic that affects both adults and children. A combination of excessive food intake and less or no physical activity is the primary cause of obesity [6, 7]. As a result, it is critical to precisely measure diet [8]. Early in the new millennium, it was noted that the rise of the chronic disease known as obesity had caused individuals to become more health-conscious. Nowadays, as a result of the development of technologies, individuals are more conscious of the foods they consume to state their hunger. This occurs as a result of obesity becoming a worldwide issue in the last few years. According to estimates from the World Health Organization (WHO), 650 million people worldwide (aged 18 and over) are obese, while there are an estimated 1.9 billion overweight people worldwide. In 2018, there were 40 million overweight children that is a worrying trend for the world [9]. The practice of keeping a food journal has become more widespread as a technique to help people monitor the number of calories they need to eat at any given moment to prevent problems of this nature [10]. Thus, the idea of calorie estimates and meal detection using a camera or smart phone has emerged, along with digital journaling. However, there is a great deal of doubt because there are several varieties of various food imageries from various world cuisines available.

The scientists gave us a number of theories after years of research into this issue. Using the recent technologies, a number of methods were created to track daily dietary consumption, where simply taking a photo of the meal was sufficient. To identify the calories present in a specific cuisine, these applications employ image processing techniques. In order to determine the number of calories ingested by a certain user, multiple algorithms for food image identification were used. The analysts used methods involving support vector machine (SVM), convolutional neural network (CNN), selective search, Gabor's filter, K-means clustering, etc. to measure food intake and gather dietary data, and they developed an intuitive and simple solution [11]. Most existing technologies are inaccurate in recognizing food items, let alone estimating total calories. To address this issue, we describe in this paper a hybrid framework for food identification and calorie estimates that combines deep learning methods such as Mask RCNN and YOLO V5.

A hybrid system for deep learning, in the context of managing dietary choices and calorie intake, combines the strengths of deep learning with other techniques or data sources to enhance its capabilities.

1.1. Practical Applications. Deep learning can be used to improve the accuracy of food recognition in images, which is particularly useful in dietary tracking apps. By combining deep learning with other technologies like computer vision, the system can better identify and log foods, making it easier for users to track their dietary choices. Deep learning can assist in real-time nutrient analysis by rapidly processing and interpreting data from food images. This helps users make immediate decisions about their food choices based on their dietary objectives. By analysing user data over time, deep learning can help predict dietary behaviour. It can identify patterns or trends in an individual's eating habits, aiding in proactive dietary management and lifestyle adjustments.

1.2. Implications. A hybrid system for deep learning in managing dietary choices and calorie intake has the potential to significantly enhance the accuracy and efficiency of dietary tracking and provide more personalized recommendations. However, it should be developed and used with considerations for data privacy and resource constraints to ensure its benefits are maximized for individuals seeking to manage their diet.

The following are the paper's primary contributions:

(i) The dataset includes the images with foods in the bowl that are collected

(ii) To segment the images of the food items, Mask RCNN is used

(iii) YOLO V5 framework is used to classify the item of food after the features are extricated

(iv) Then, the dimension of each food item is determined. Based on the estimated dimension, food item's volume is evaluated

(v) Then, the calorie in intake food is estimated using the food item's volume

This complete article is structured as follows: the study based on food recognition and calorie estimation that has been done by various researchers is compiled in Section 2. Then, Section 3 provides specifics on the proposed methodology for our research. Section 4 illustrates and explains the consequences and results of several strategies. The conclusions and recommendations for further research are presented in Section 5.

2. Related Works

Kumar et al. [1] introduced a technique for food item detection and calorie prediction that makes use of SVM and enhanced MLP models. This work was done for a particular food item using different preprocessing approaches, segmentation, and analysis of extraction. For recognition, the collected features were input into SVM and MLP classifiers. By calculating food item calorie values automatically, this application assisted diabetes patients in maintaining healthy diet and BMI. Furthermore, when compared to the SVM, MLP was capable of obtaining calorie values that were close to the real sample values. As a result, the proposed technique obtained great accuracy and efficiency when recognizing food items. However, this approach does not deal with complicated food products that are studied from several angles, and it has to be improved.

Turmchokksam and Chamnongthai [12] presented a method for ingredient-based food calorie measurement integrating thermal data and nutritional knowledge. The components of the selected meal, along with their nutritional data, brightness pattern, and thermal images, were then retrieved via the database. The image was separated into candidate ingredient borders, and using fuzzy logic, each candidate ingredient boundary was then classified as an ingredient based on the patterns and intensities of its heat. The categorized items from each border were finally estimated for total calories determined by area ratio and nutritional data. Despite the fact that their technology efficiently tracked the number of calories in each meal, several people considered the component cost and system scale to be disadvantages.

Shen et al. [13] presented a machine learning-based strategy for food recognition and nutrient prediction. This technique made use of a brand-new system built on transfer learning that accurately classified food images and evaluated food qualities automatically. This strategy offered a mechanism for improving the Inception V-4 and V-3 models to identify the food items while also quantifying the attributes of the food by the attribute estimation model. Data augmentation and other similar techniques were used to improve the results. The accuracy of the suggested method for classifying data and extracting attributes was 85%, but more work needs to be done to improve the system's usability and accuracy.

A compact and effective convolutional neural network architecture for Chinese food recognition was described by Teng et al. in [14]. The goal of this network architecture was to model and implement a processing pipeline that was similar to the bag-of-features (BoF) technique. The two phases of the proposed CNN architecture were feature extraction and feature combining. A conventional BoF method was used to define every region of the food image, highlighting the particular peculiarities of the input image with specified descriptors. Through the application of learning algorithms, local characteristics of the food image were utilized to create a visual dictionary. CNN's inherent unified optimization gave them the best results for identifying Chinese food, but the downsides of slower classification speeds and a tiny increase in network complexity came at the expense of accuracy.

Ma et al. [15] demonstrated a method for estimating Chinese market food nutrients using image-based deep learning algorithms. The first step involved collecting several food images with commentary through the Internet. The second phase was carrying out a conventional series of image preparation techniques, like pixel value normalization and image scaling, to lessen the influence from resolution irregularities and uneven illumination. The last step proceeded to use the training set to train CNN models for image classification, after which the model conclusions were assessed using the validation set. However, further developments are needed to increase the estimation accuracy of some nutrients.

A novel method for estimating food volume utilizing a wellness model for calculating calories was proposed by Kadam et al. [16]. For accurate segmentation of food images with uniform and uneven forms on many food plates, this study applied mask-based RCNN. As a volume estimator model, the RCNN-based food segmentation was used. It was created by tweaking the previously trained ResNet model, which was trained across a dataset of eight different classes of photos of foods in all forms. Furthermore, utilizing reference-based estimations, the pixel to size converter was used to validate the size estimates for the food items and containers. The accuracy of unstructured food items, which was in the 90%-91% range, needs to be further improved.

3. Hybrid Deep Learning Algorithm-Based Food Recognition and Calorie Estimation

3.1. Overview. Figure 1 depicts an overview of the proposed scheme. In this study, Indian FoodNet-30 dataset is used to analyze the proposed scheme. The dataset includes the images with foods in the bowl. Then, the Mask RCNN is used to segment the images of the food items. From the segmented images, features are extracted, and the food item is classified using YOLO framework. After recognizing the food items, dimension of each food item is determined. Based on the estimated dimension, volume of the food item is calculated. Then, the calorie is estimated using the volume of the food item. Overview of the proposed scheme is given in Figure 1.

3.1.1. Dataset Description. Ritu Agarwal, Nikunj Bansal, Tanmay Sarkar, Tanupriya Choudhury, and Neelu Jyothi Ahuja created Indian FoodNet-30 with the purpose of developing an Indian food detection model which includes almost 5500 images of 30 prominent Indian foods. The dataset includes the following classes: aloo gobi, aloo masala, bhatura, bhindi masala, biryani, chai, chole, coconut chutney, dal, dosa, dum aloo, fish curry, ghevar, green chutney, gulab jamun, idli, jalebi, kebab, kheer, kulfi, lassi, mutton curry, onion pakoda, palak paneer, poha, rajma curry, ras malai, samosa, shahi paneer, and white rice. Besides, each image in the dataset is resized to 640 × 640. 80% of the data in this study is utilized to train the model. The rest of the model is employed to put the trained model to the test. The link for the dataset is https://universe.roboflow.com/indianfoodnet/indianfoodnet.

3.1.2. Segmentation Using Mask RCNN. After the preparation of the dataset, the images are sent into Mask RCNN for segmentation. The Mask RCNN is a deep learning-driven technique for detecting objects and segmenting instances. Instance segmentation gives different masks for different instances of the same class of objects. The system design is separated into two sections; the first is RPN (region proposal network), and it is in charge of producing probabilities that the food items are located within the image with the pretrained network ResNet.

Mask RCNN's architecture is depending on the Faster RCNN object detection algorithm that has two major parts, the RPN and Fast R-CNN network [17]. Mask RCNN includes an additional branch, for instance, segmentation in addition to these components. Figure 2 represents the architecture of Mask RCNN design.

The 3rd branch exists on the apex of the current Faster RCNN architecture; it shares the same feature extraction backbone as the other two branches. The convolutional feature map is slid over by a tiny network produced by the backbone network by the first component, the RPN, in order to generate region proposals.

The second component, the Fast RCNN network, takes these region proposals as input and performs classification and bounding box regression to generate object detections. The third branch, for instance, segmentation, takes the same region proposals as input and generates binary masks that indicate which pixels in each proposal belong to the object of interest. This branch is built using a fully convolutional network (FCN) that shares weights with the other two branches [18]. The components in Mask RCNN are explained as follows.

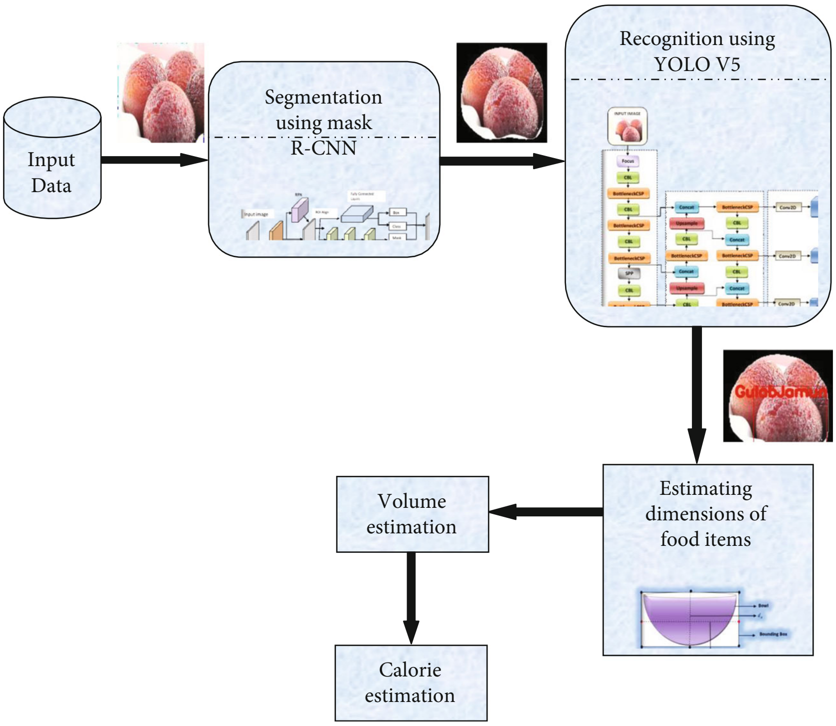

Input Data
Segmentation using mask R-CNN
Recognition using YOLO V5
GulabJamun
Estimating dimensions of food items
Volume estimation
Calorie estimation

Figure 1: The overview of the proposed scheme.

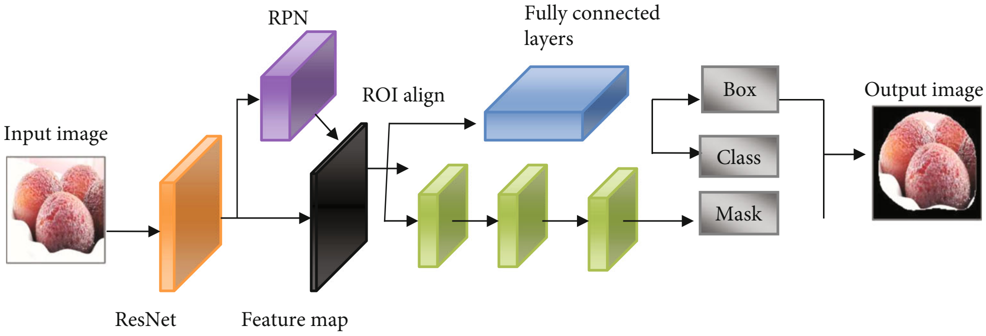


Figure 2: Architecture of Mask RCNN.

(1) ResNet-FPN. ResNet-FPN is a backbone architecture used for feature extraction in Mask RCNN. A deep CNN called ResNet has proven to be quite successful in classifying images. In ResNet-FPN, FPN architecture is added on top of the ResNet backbone to create a more effective feature extractor. The FPN component allows for multiscale feature maps to be generated from the input image, which can improve object detection accuracy [19].

(2) RPN. The RPN network receives the features that were extracted using ResNet101+FPN as input to produce ROIs. RPN may predict the front and back of an image if the fold length and breadth ratios are different. To quickly create candidate areas, place the image box in this case on the network and describe the border-box in the anticipated feature image. A 3×3 convolutional layer serves to scan the image and produce the relating anchors, which are dispersed across the image in various sizes.

A network scales itself in accordance with the input pictures and has a set of anchor boxes with predefined positions. The bounding boxes and ground truth classes are each given a unique anchor. To recognize defects of various sizes and shapes, default bounding boxes come in a range of sizes as well as aspect ratios. These boxes overlap one another numerous times, which makes it easier to determine the greatest confidence score for detecting several ROIs. It is known as an intersection over union factor (IoU), and it is calculated by using

$$\text{IoU} = \frac{\text{area of overlap}}{\text{area of union}}. \tag{1}$$

(3) ROI Align. ROI Align takes a set of rectangular region proposals as input and extracts feature from the feature map that correspond to each proposal. High pixel accuracy and the ability to determine whether each branch is a part

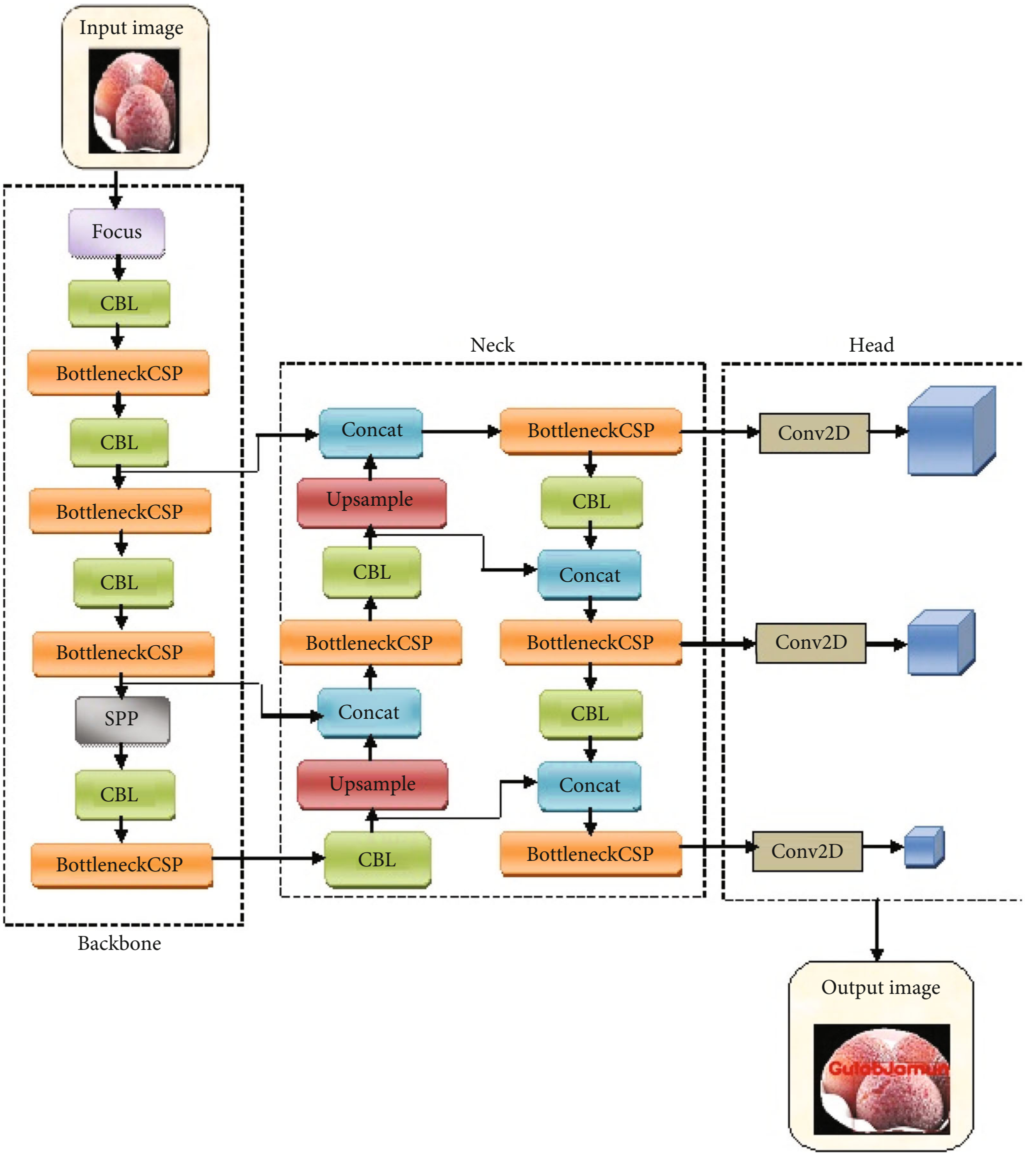


Figure 3: YOLO V5 network structure.

of the same pixel target are requirements for an RCNN network that enables mask branch detection of pixel targets. Following the pooling and convolution of the original image, the image's size is changed. Segmentation is then carried out. The proper segmentation output cannot be produced by the direct pixel-level segmentation technique. Therefore, Mask RCNN, an improved version of Faster RCNN, is suggested in this paper. Additionally, the pooling layer of the CNN is changed to ROI Align. This approach employs linear interpolation to preserve the spatial details of the feature map. ROI Align is a neural network layer used in object detection as well as instance segmentation algorithms such as Mask RCNN.

(4) FCN. FCN is in charge of the mask segmentation conducted on every region of interest (ROI). FCN predicts using a pixel-by-pixel technique. FCN employs the per-pixel softmax loss function.

Loss function

The multitask loss function (Lf) is specified by the Mask RCNN algorithm for every region of interest (ROI) collected during training which is given as

$$\mathrm{Lf} = \mathrm{Lf}_{\mathrm{cls}} + \mathrm{Lf}_{\mathrm{box}} + \mathrm{Lf}_{\mathrm{mask}}. \tag{2}$$

Here, classification error is denoted as $\mathrm{Lf}_{\mathrm{cls}}$, detection error is denoted as $\mathrm{Lf}_{\mathrm{box}}$, and segmentation error is denoted as $\mathrm{Lf}_{\mathrm{mask}}$.

In Mask RCNN, $\mathrm{Lf}_{\mathrm{cls}}$ and $\mathrm{Lf}_{\mathrm{box}}$ are represented as

$$\mathrm{Lf}' = \frac{1}{W_{\mathrm{cls}}} \sum_i \mathrm{Lf}_{\mathrm{cls}}\left(p_i, p_i^*\right) + \lambda \frac{1}{W_{\mathrm{reg}}} \sum_i p_i^* \mathrm{Lf}_{\mathrm{reg}}\left(v_i, v_i^*\right), \tag{3}$$

where p_i denotes the i^{th} target's projected probability of being on the anchor point. Using the sign of the anchor point sample, p_i^* is calculated. If the sample from the anchor point is positive, then p_i^* is 1, or else 0.

The v_i and v_i^* are vectors made up of four translational and scaling factors, which, in turn, reflect how much the positive sample anchor point has changed in relation to the prediction area as well as the label area, respectively. To maintain equilibrium, the weights W_{cls}, W_{reg}, and λ regulate the two losses.

Regression loss and classification loss are given as

$$\begin{aligned} \mathrm{Lf}_{\mathrm{reg}}(v_i, v_i^*) &= \mathrm{smooth}_{\mathrm{Lf}}(x)(v_i, v_i^*), \\ \mathrm{Lf}_{\mathrm{cls}}(p_i, p_i^*) &= -\log\left[p_i, p_i^* + (1-p_i^*)(1-p_i)\right]. \end{aligned} \tag{4}$$

The robust loss $\mathrm{smooth}_{\mathrm{Lf}}(x)$ is defined by the corrected frame's x translation at the anchor point on the horizontal axis. It is described as

$$\mathrm{smooth}_{\mathrm{Lf}}(x) = \begin{cases} 0.5x^2 & |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{5}$$

An average binary cross-entropy function called $\mathrm{Lf}_{\mathrm{mask}}$ in Mask RCNN is used to characterize the loss of semantic segmentation branches [20]. After processing, the input feature map in the mask branch is output into a $k \times m \times m$ format, where k and m determine the feature map's size and scale, respectively. The resulting feature map's pixel-by-pixel sigmoid computation yields the relative entropy, and the average entropy error is $\mathrm{Lf}_{\mathrm{mask}}$.

As a result, the suggested Mask RCNN effectively segments the location of the food objects inside the image.

3.2. Food Recognition Using YOLO V5. After segmentation, the segmented images are sent into the YOLO V5 classifier for further classification. Features are retrieved from those segmented images, and the food items are then identified using the YOLO V5 network. Redmon J. presented the YOLO object identification technique as the first single-stage object detection system. After feeding the full image into the network, the essential premise of YOLO is to immediately reverse the positions and types of the bounding boxes on the output layer, which is equivalent to Faster RCNN. The candidate box extraction phase from the 2-stage strategy is removed by combining the classification issues as well as bounding box onto a regression problem. In comparison to earlier iterations of the YOLO network, version 5 has significantly increased detection accuracy while maintaining the benefit of quick detection speed, which helps to offset the disadvantages of the one-stage network's poor detection accuracy. Figure 3 depicts the network architecture of YOLO V5. The YOLO is made up of three major components:

Backbone: convolutional neural network that gathers and produces image characteristics based on image resolution

Neck: group of network layers that aggregate and blend picture features before passing them on to the prediction layer

Head: in addition to producing bounding boxes and predicting categories, it can predict image attributes. The level

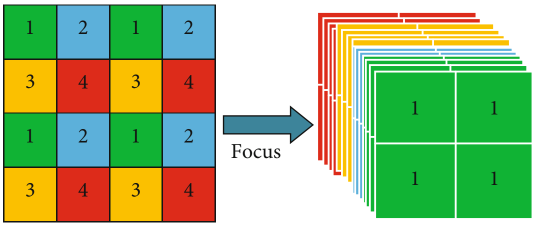


Figure 4: Processing flow of focus model.

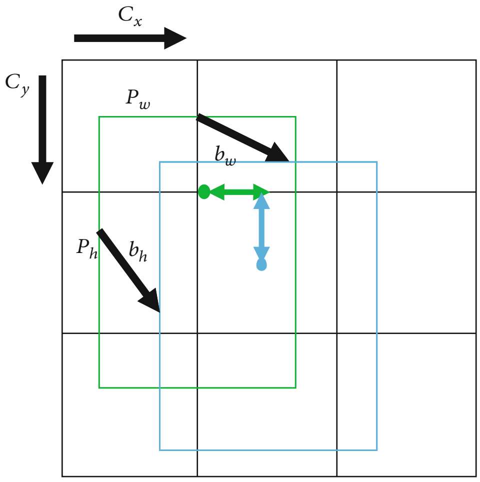


Figure 5: Decoding of prediction bounding box in YOLO V5.

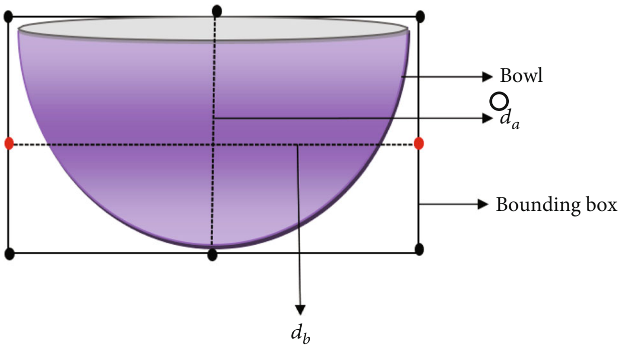


Figure 6: Bowl's dimensions marking using the pixel technique.

Table 1: Simulation parameters.

Batch_size	2
Learning rate	0.001
Weight_decay	1e − 3
Optimizer	Adam
Epochs	25
Split_ratio	0.2
Device	CPU
Dataset for training	80%
Dataset for testing	20%

Table 2: Segmented and detected images of the proposed and existing approaches.

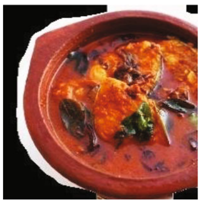
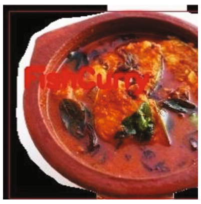
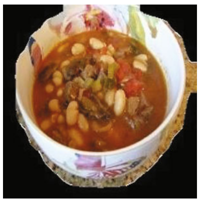
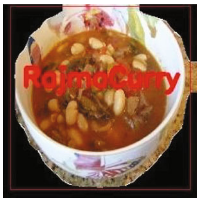
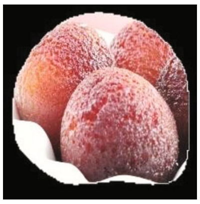
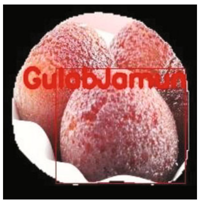
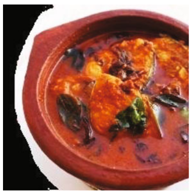
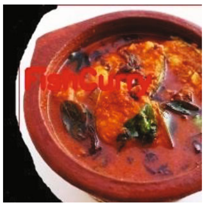
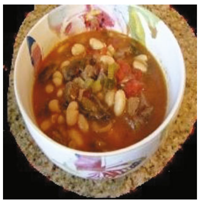
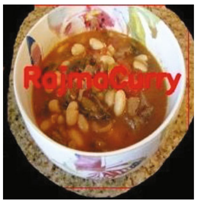
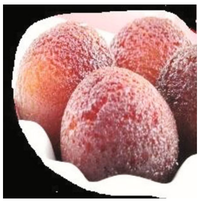
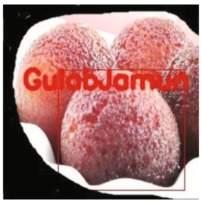
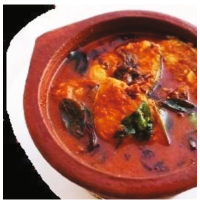
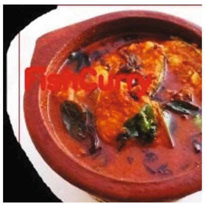
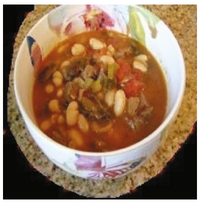
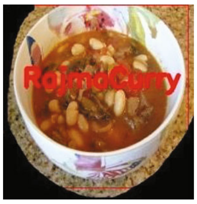
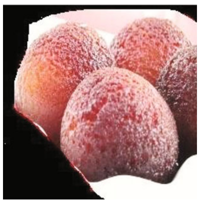
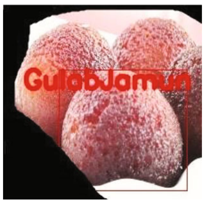
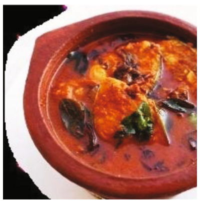
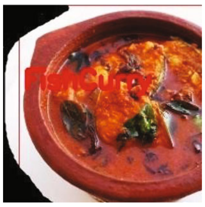
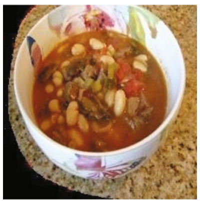
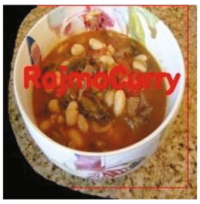
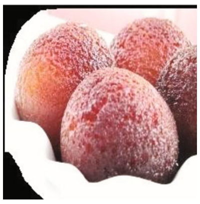
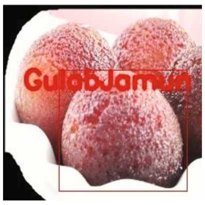
Methods	Input image	Segmented image	Detected image
Mask RCNN+YOLO V5			
			
			
Mask RCNN			
			
			

Table 2: Continued.

Methods	Input image	Segmented image	Detected image
YOLO			
			
			
CNN			
			
			

of confidence reveals how accurately a categorization was made in a given circumstance

The three parts of the YOLO V5 framework are the backbone, neck, and output. The focus component in the backbone is where the input image with the specified resolution goes. The CBL is a simple convolution module. Conv2D +Batch Normal+Leaky RELU are symbols for a CBL module. The Bottleneck CSP module basically performs feature extraction upon the feature map in order to extract rich data from the image. Its parameter amount takes up the majority of the variable quantity across the whole system. The SPP

Table 3: Volume and calorie estimations.

S. no	Food items	Estimated volume	Estimated calories
1	Fish curry	47944 cubic cm	217 calories
2	Rajma curry	28245 cubic cm	207 calories
3	Gulab jamun	40015 cubic cm	300 calories

Table 4: Comparative analysis of the existing approaches with the proposed method.

Evaluation metrics	Mask RCNN+YOLO V5	Mask RCNN	YOLO	CNN
Accuracy	0.97125	0.96225	0.95125	0.92125
Sensitivity	0.96754	0.95354	0.94754	0.93754
Specificity	0.96452	0.95152	0.94252	0.93452
Precision	0.96415	0.95015	0.94115	0.93415
F-score	0.96584	0.95184	0.94433	0.93584
MCC	0.96241	0.95041	0.94341	0.93241
NPV	0.94592	0.93892	0.92692	0.91592
FPR	0.0154	0.0214	0.0294	0.0354
FDR	0.0148	0.0198	0.0238	0.0348
FNR	0.0161	0.0251	0.0321	0.0461

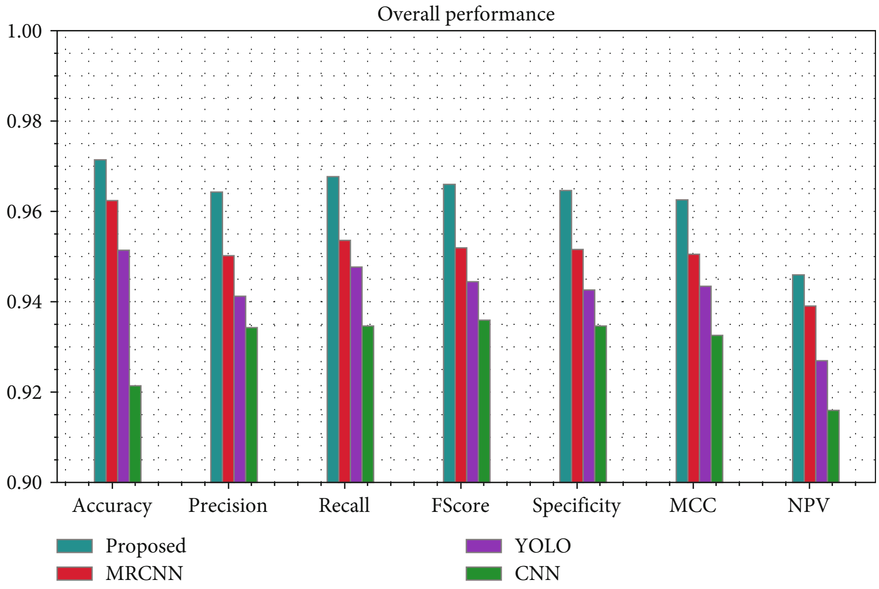


Figure 7: Analysis of already existing approaches with the proposed method.

module adds functionality of different kinds and largely increases the network's receptive area.

YOLO V5 also incorporates a bottom-up feature arrangement built upon the FPN framework. In this combined operation, the FPN layer communicates strong semantic information from top to bottom while the feature pyramid sends solid location characteristics in the bottom up. Display the item coordinates and categorization results at the bottom of the figure.

3.2.1. Input. The input consists of three components: automated adaption anchor frame, picture size processing, and data improvement. By improving mosaic data, the classic YOLO V5 increases the detection of small objects by cropping, setting up, and fusing the input. Prior to feeding the input image into the model for analysis, as part of dataset training, the size of the input image is adjusted to make it uniform.

3.2.2. Backbone. The focus structure and CSP structure make up the backbone network. Prior to entering the backbone

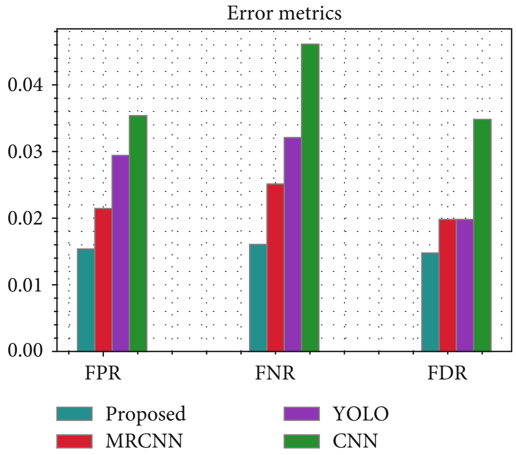
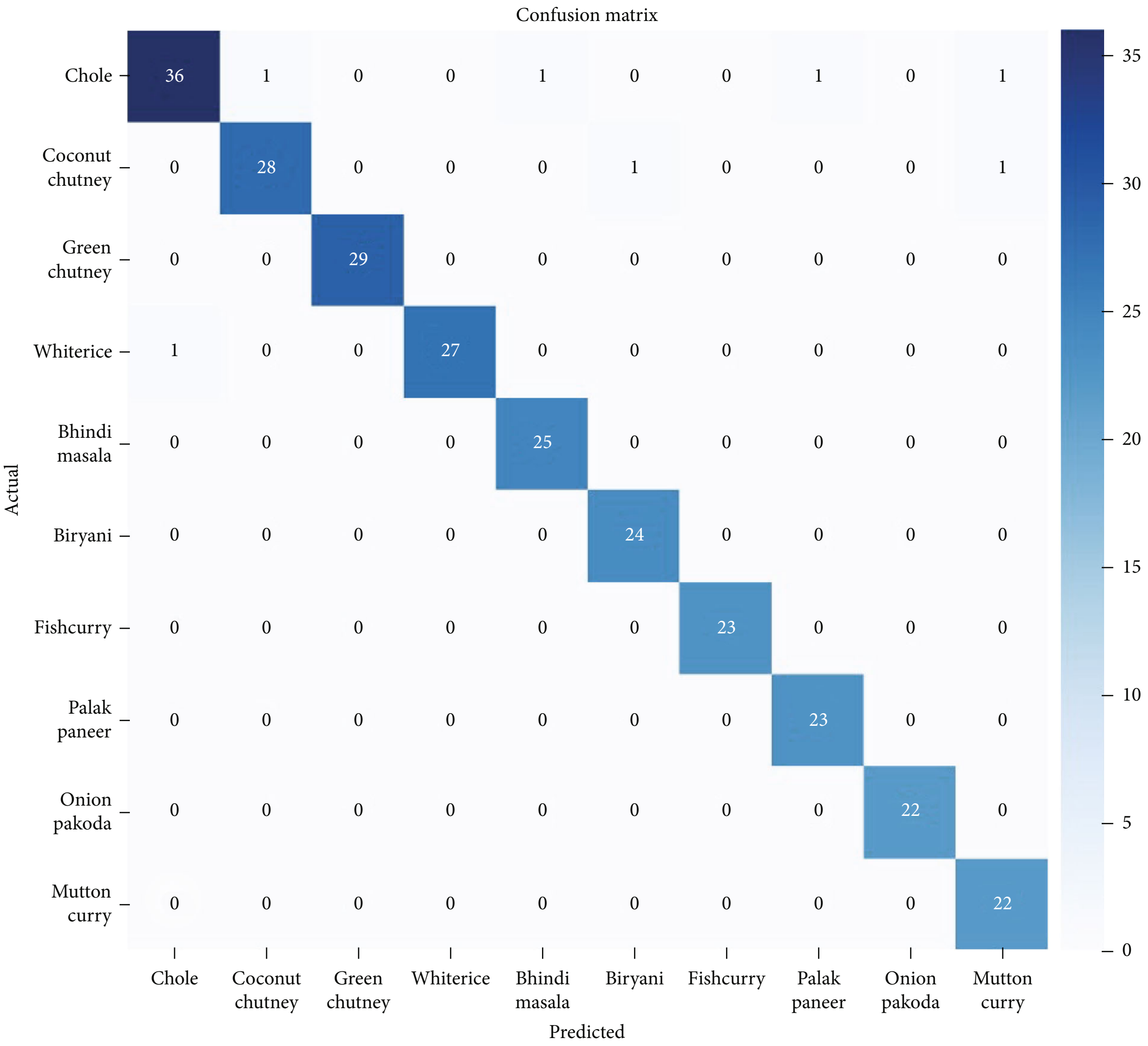


Figure 8: Comparative analysis of error metrics of various approaches.



	Chole	Coconut chutney	Green chutney	Whiterice	Bhindi masala	Biryani	Fishcurry	Palak paneer	Onion pakoda	Mutton curry
Chole	36	1	0	0	1	0	0	1	0	1
Coconut chutney	0	28	0	0	0	1	0	0	0	1
Green chutney	0	0	29	0	0	0	0	0	0	0
Whiterice	1	0	0	27	0	0	0	0	0	0
Bhindi masala	0	0	0	0	25	0	0	0	0	0
Biryani	0	0	0	0	0	24	0	0	0	0
Fishcurry	0	0	0	0	0	0	23	0	0	0
Palak paneer	0	0	0	0	0	0	0	23	0	0
Onion pakoda	0	0	0	0	0	0	0	0	22	0
Mutton curry	0	0	0	0	0	0	0	0	0	22

Figure 9: Confusion matrix of the proposed method.

network, the image is sliced by the focus structure. The original image is divided, as seen in Figure 3. Once the feature map of half of the image resolution is reached, the convolutional operation is used to create the feature map. Without using any settings, the focus can reduce the input sizes while keeping as much of the original image data as is practical. Two 1*1 convolutions are used to transition the input feature into the CSP structure. It helps with overcoming computing bottlenecks, lowering memory costs, and enhancing CNNs' capacity for learning.

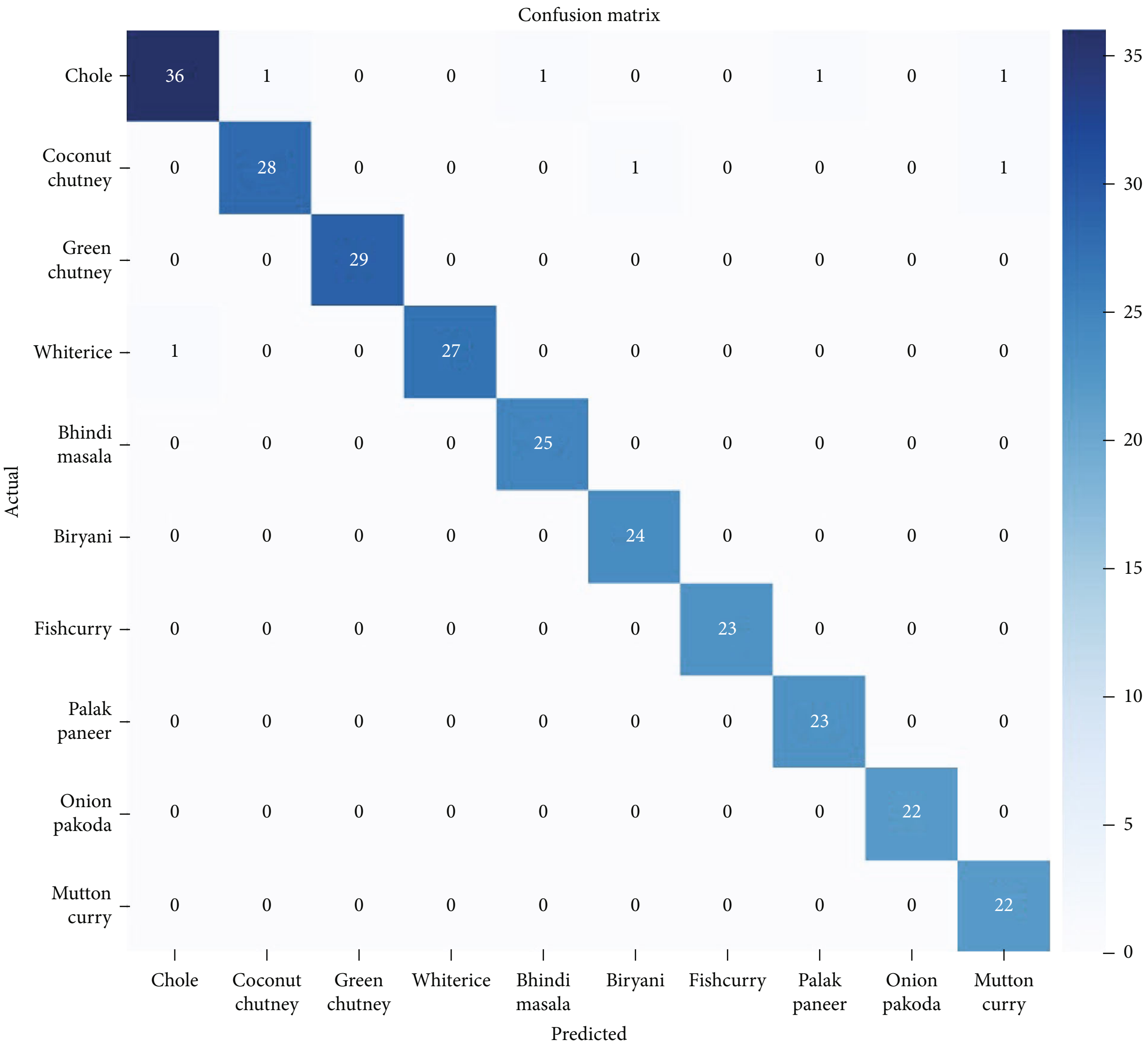
3.2.3. Neck. The prediction layer receives the combined picture features from the neck portion, a network layer. The FPN+PAN is used in YOLO V5 neck portion. The FPN communicates and merges the top-to-bottom high-level feature information to generate a feature map for forecasting. The underlying pyramid, known as PAN, communicates powerful positional features via bottom to top.

3.2.4. Output Terminal. This prediction layer creates the bounding box to forecast both the picture's characteristics and its categorization. Its loss function is CIoU loss that YOLO V5 uses for bounding box. CIoU NMS performs better than classic nonmaximum suppression (NMS) at detecting overlapping objects. Figure 4 shows the focus model of the processing of flow.

The YOLO algorithm's connecting ideas are carried over to YOLO V5. That is the 3-channel RGB color picture's length and breadth following the first image preparation. As the three scales of big, medium, and micro go to the network, the output dimension of the detecting layer is

$$S \times S \times n_a \times \left(t_x + t_y + t_w + t_h + t_o + n_c\right), \tag{6}$$

$S \times S$ is the divided mesh's number. The symbol n_a designates the number of predetermined preceding boxes on every scale. There are a number of things that need to be



	Chole	Coconut chutney	Green chutney	Whiterice	Bhindi masala	Biryani	Fishcurry	Palak paneer	Onion pakoda	Mutton curry
Chole	36	1	0	0	1	0	0	1	0	1
Coconut chutney	0	28	0	0	0	1	0	0	0	1
Green chutney	0	0	29	0	0	0	0	0	0	0
Whiterice	1	0	0	27	0	0	0	0	0	0
Bhindi masala	0	0	0	0	25	0	0	0	0	0
Biryani	0	0	0	0	0	24	0	0	0	0
Fishcurry	0	0	0	0	0	0	23	0	0	0
Palak paneer	0	0	0	0	0	0	0	23	0	0
Onion pakoda	0	0	0	0	0	0	0	0	22	0
Mutton curry	0	0	0	0	0	0	0	0	0	22

FIGURE 10: Confusion matrix of Mask RCNN.

predicted (n_c). An example of a network structure is the large scale, $S = 20$, $n_a = 3$, and $n_c = 3$, and the output dimension of the detecting layer is 9600.

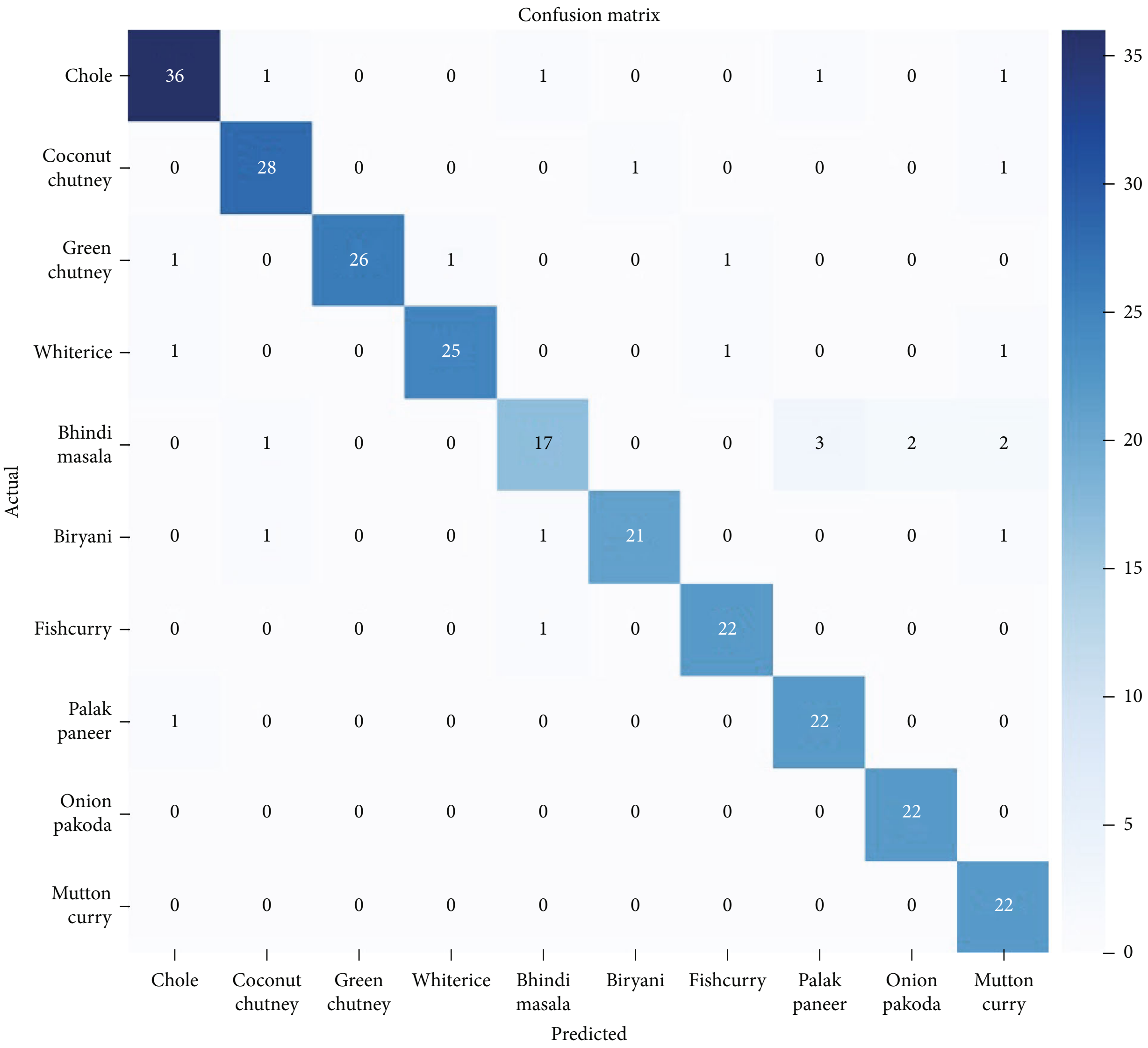
The parameters linked to bounding boxes are t_x, t_y, t_w, and t_h. The bounding box's level of confidence is t_o. Then, the confidence of category i is t_{ci}. These parameters need the following equations to be decoded in order to get the final prediction box.

$$\begin{aligned}
a_x &= 2\sigma(t_x) - 0.5 + c_x, \\
a_y &= 2\sigma\left(t_y\right) - 0.5 + c_y, \\
a_w &= p_w(2\sigma(t_w))^2, \\
a_h &= p_h(2\sigma(t_h))^2, \\
\text{Score}_i &= \text{confidence} \times \Pr(\text{class}_i) = \sigma(t_o)\sigma(t_{ci}).
\end{aligned} \tag{7}$$

Figure 5 depicts the ground truth box as the blue box. The label bounding box's centre point, coordinates, width, and height are given as a_x, a_y, a_w, and a_h, respectively. The distances c_x and c_y are measured from the upper left corner of the grid to the centre of the label bounding box. The anchor box is the red one. The previous frame's width and height are p_w and p_h, respectively.

Positive-negative (foreground-background) sample imbalance is a frequent problem that most adversely affects algorithm performance when training on an object detection task. Throughout training, YOLO V5 sends the label box to 3 anchors simultaneously, so it is comparable to double the positive sample size. The discrepancy among each of the positive and negative samples is slightly corrected during the target detection algorithm's training.

Formula (8), which calculates the loss function, is used. L_{CIoU} loss is employed as the loss function of the bounding box.



Actual \ Predicted	Chole	Coconut chutney	Green chutney	Whiterice	Bhindi masala	Biryani	Fishcurry	Palak paneer	Onion pakoda	Mutton curry
Chole	36	1	0	0	1	0	0	1	0	1
Coconut chutney	0	28	0	0	0	1	0	0	0	1
Green chutney	1	0	26	1	0	0	1	0	0	0
Whiterice	1	0	0	25	0	0	1	0	0	1
Bhindi masala	0	1	0	0	17	0	0	3	2	2
Biryani	0	1	0	0	1	21	0	0	0	1
Fishcurry	0	0	0	0	1	0	22	0	0	0
Palak paneer	1	0	0	0	0	0	0	22	0	0
Onion pakoda	0	0	0	0	0	0	0	0	22	0
Mutton curry	0	0	0	0	0	0	0	0	0	22

FIGURE 11: Confusion matrix of YOLO.

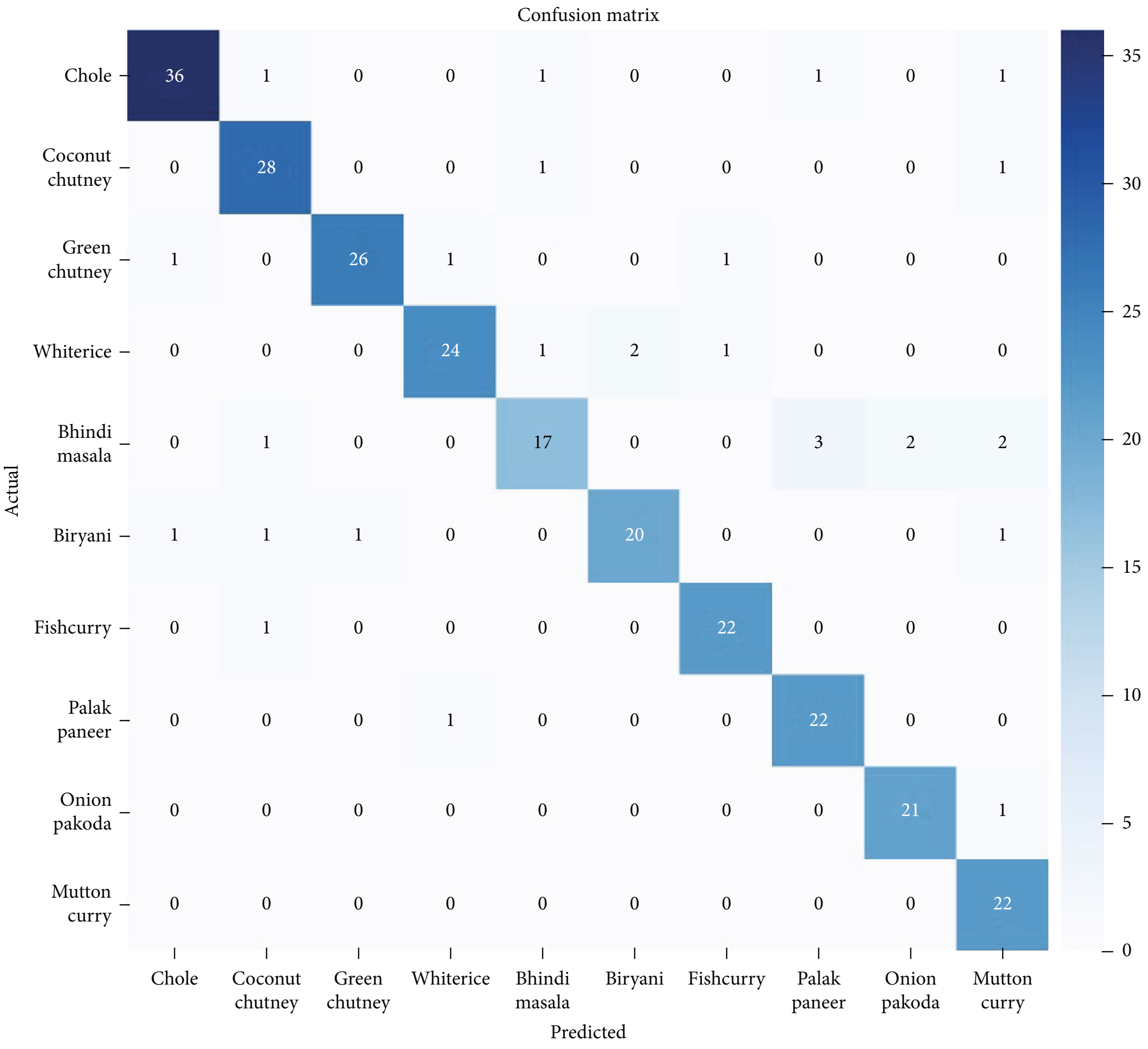
$$L_{\text{total}} = \sum_{i}^{N}\left(\lambda_1 L_{\text{box}} + \lambda_2 L_{\text{obj}} + \lambda_3 L_{\text{cls}}\right) = \sum_{i}^{N}\left(\lambda_1 \sum_{j}^{B_i} L_{\text{CIoU}_j} + \lambda_2 \sum_{j}^{S_i\times S_i} l_{\text{objj}} + \lambda_3 \sum_{j}^{B_i} l_{\text{cls}_j}\right). \tag{8}$$

There are N layers of detection. The box in front of it has B labels, which is the desired number. This scale $S \times S$ is made up of number of meshes. L_{box} is the bounding box that each target's regression loss calculation uses. The target object's loss L_{obj} is used to calculate each mesh. Each target was determined using the categorized loss, which is L_{cls}. These three losses have the following weights λ_1, λ_2, and λ_3.

3.3. Volume Estimation. The volume of the item will be measured following that the ROI has been clearly designated and the object has been classified. Here, a coin is utilized as a symbol. As a result, two bounding boxes are marked on the coin and the object for volumetric estimation. The reference bowl is shown in Figure 6, along with the bowl's diameter (d_a) as well as height (d_b) dimensions.

The volume of irregular food items is estimated by the pixel-per-meter approach. This approach determines the reference object's width in pixels. Using this ratio, the object's height along with breadth is further calculated. The estimated volume (Vol_E) of the bowl is then determined using these measurements in Eq. (11) and is essentially the quantity of food items with bowl-like shapes. Equation (11) is obtained from the traditional equation for the hemispherical bowl found in

$$\text{Volume}_{\text{Bowl}} = \frac{\pi}{6}\left(3(r_{\text{Bowl}})^2 + (h_{\text{Bowl}})^2\right)h_{\text{Bowl}}, \tag{9}$$



Actual \ Predicted	Chole	Coconut chutney	Green chutney	Whiterice	Bhindi masala	Biryani	Fishcurry	Palak paneer	Onion pakoda	Mutton curry
Chole	36	1	0	0	1	0	0	1	0	1
Coconut chutney	0	28	0	0	1	0	0	0	0	1
Green chutney	1	0	26	1	0	0	1	0	0	0
Whiterice	0	0	0	24	1	2	1	0	0	0
Bhindi masala	0	1	0	0	17	0	0	3	2	2
Biryani	1	1	1	0	0	20	0	0	0	1
Fishcurry	0	1	0	0	0	0	22	0	0	0
Palak paneer	0	0	0	1	0	0	0	22	0	0
Onion pakoda	0	0	0	0	0	0	0	0	21	1
Mutton curry	0	0	0	0	0	0	0	0	0	22

FIGURE 12: Confusion matrix of CNN.

$$h_{\mathrm{Bowl}} = \frac{d_{\mathrm{a}}}{\mathrm{pPM}}, \qquad r_{\mathrm{Bowl}} = \frac{1}{2}\frac{d_{\mathrm{b}}}{\mathrm{pPM}}, \tag{10}$$

$$\mathrm{Vol}_{\mathrm{E}} = \frac{\pi}{6}\frac{d_{\mathrm{a}}}{\mathrm{pPM}}\left(\frac{3}{4}\left(\frac{d_{\mathrm{b}}}{\mathrm{pPM}}\right)^2\left(\frac{d_{\mathrm{a}}}{\mathrm{pPM}}\right)^2\right), \tag{11}$$

$$\mathrm{pPM} = \frac{\text{object-width}}{\text{know-width}}. \tag{12}$$

Here, d_{a} and d_{b} are the Euclidean distances among 2 points on the opposing sides of the items, object-width refers to the width of the bowl, pPM stands for pixel-per-metric ratio, and know-width refers to the size of the reference object (coin), which is equal to 0.9 inches.

3.4. Calorie Estimation. Calories can be computed when the volume has been determined. For both cooked and uncooked meals, there are standard calorie mapping charts available [21]. Calories per gram are shown on the graph. Volume is transformed into grams and linked to the calorie table in order to determine the exact number of calories in food item that our system has identified. Here, C_{E} is the calorific equivalent, and total calorie ($\mathrm{Calorie}_{\mathrm{total}}$) is calculated by using the below formula:

$$\mathrm{Calorie}_{\mathrm{total}} = \mathrm{volume} * C_{\mathrm{E}}. \tag{13}$$

Thus, the calorie is estimated using the volume of the food item.

4. Results and Discussion

This section investigates the performance of the model which is proposed here. The proposed technique is implemented

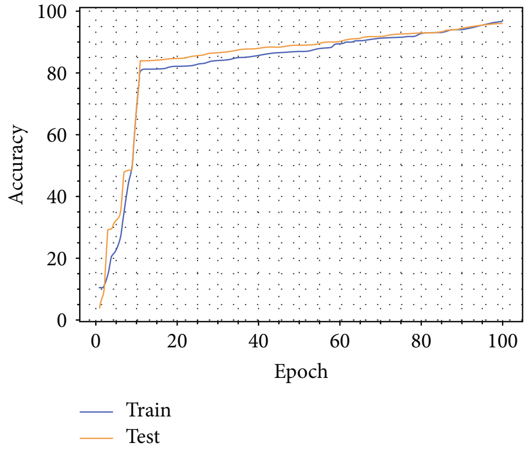


Figure 13: Accuracy-epoch graph.

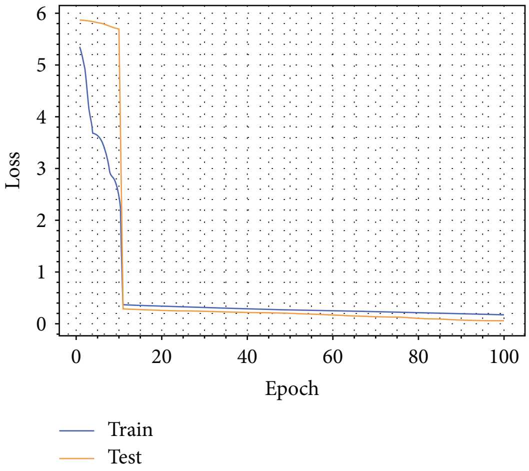


Figure 14: Loss-epoch graph.

using Python. The system specifications for the implementation include a 1.6 GHz Intel Core i5 Processor running Windows and 4 GB memory of RAM. The testing results of three separate images are segmented using Mask RCNN and classified using YOLO V5, and this method achieved maximum accuracy, sensitivity, specificity, precision, recall, FDR (false discovery rate), *F*-score, MCC, FPR (false positive rate), NPV, and FNR (false negative rate) because of these two methods. Furthermore, the effectiveness of the proposed YOLO V5 is compared and analysed with various segmentation and classification models such as Mask RCNN, YOLO, and CNN. Simulation parameters are depicted in Table 1. Table 2 depicts the segmented and identified images from both proposed and existing approaches. Table 3 shows a sample of estimated volume on detected image, and on the basis of volume, calorie is evaluated.

4.1. Performance Analysis. The performance of the proposed method is analysed in terms of accuracy, sensitivity, specificity, precision, recall, *F*-score, MCC, NPV, FPR, FDR, and FNR. Moreover, the efficiency of the proposed Mask RCNN+YOLO V5 is compared and analysed with other classification models such as Mask RCNN, YOLO, and CNN. The following section describes the performance analysis of the three different images in terms of different metrics using various classification models.

In Table 4, the evaluation metrics of the proposed Mask RCNN+YOLO V5 are compared with other classification models. Among various classification models, Mask RCNN +YOLO V5 attained maximum accuracy of 0.97125 when compared to Mask RCNN, YOLO, and CNN. On average, the accuracy is 0.96225 for Mask RCNN, 0.92125 for CNN, and 0.95125 for YOLO, which is lower compared to Mask RCNN+YOLO V5. The sensitivity and specificity of our proposed method are 0.96754 and 0.96452, respectively, which are higher than other models. Our proposed method attained 0.96241 and 0.94592 in terms of MCC and NPV. The precision and *F*-score of our proposed method are 0.96415 and 0.96584, respectively, which are higher than the other three models. Figure 7 shows the comparative analysis of the proposed and already existing approaches.

The error metrics of the proposed method with the existing methods are compared in Table 1. It shows that our proposed method attained very less error value when compared with other existing methods, i.e., 0.0154 for FPR, 0.0148 for FDR, and 0.0161 for FNR. Figure 8 represents the comparative analysis of error metrics of various approaches.

Figure 9 depicts the proposed approach's confusion matrix, as well as the confusion matrix for existing approaches such as the following: Figure 10 depicts Mask RCNN, Figure 11 depicts YOLO, and Figure 12 depicts CNN.

The accuracy-epoch graph is shown in Figure 13 and demonstrates how accuracy grows in parallel with increasing epoch. Figure 14 is a graph of the relationship between loss and epoch, and it demonstrates how, as epoch grows, loss value lowers.

5. Conclusion

In this study, a hybrid framework using deep learning algorithms to estimate the calorie content of food items has been proposed. Initially, the images were segmented using Mask RCNN. From the segmented images, features were extracted, and the food item is classified using YOLO V5 framework. After recognizing the food items, dimension of each food item was determined. Based on the estimated dimension, food item's volume was calculated. Then, the calorie was estimated using the food item's volume. The aforementioned models, which were developed using images of food from the dataset, have successfully identified and calculated the caloric content of food items with a high degree of accuracy and less errors. The efficiency of the proposed approach has been evaluated and compared with other classification models such as CNN, YOLO, and Mask RCNN using a variety of criteria, accuracy, NPV, sensitivity, specificity, FDR, precision, FPR, recall, *F*-score, MCC, and FNR. From those evaluation metrics, the proposed method was found to be very high with 97.12% accuracy compared to other models.

5.1. Limitations and Future Scope. One significant challenge is the variability in food types, shapes, sizes, and presentation.

Deep learning models may struggle to accurately estimate the volume of irregularly shaped or heavily processed foods. Creating accurate ground truth data for training deep learning models can be time-consuming and expensive. Accurate volume labels for a wide range of foods are often challenging to obtain. Deep learning models may not always generalize well to unseen foods or variations in presentation, which limits their real-world applicability. Variations in lighting, background, and the context in which the food is presented can impact the accuracy of deep learning models in estimating food volume. Addressing these issues may require ongoing research, data collection efforts, and the development of more robust deep learning models to make food volume estimation more accurate and practical in real-world applications.

Data Availability

Data is available at https://universe.roboflow.com/indianfoodnet/indianfoodnet.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This is to acknowledge the University of Petroleum & Energy Studies, Dehradun (UPES), for giving an opportunity to provide an environment and work as a research scholar.

References

[1] R. D. Kumar, E. G. Julie, Y. H. Robinson, S. Vimal, and S. Seo, "Recognition of food type and calorie estimation using neural network," *The Journal of Supercomputing*, vol. 77, no. 8, pp. 8172–8193, 2021.

[2] M. A. Subhi, S. H. Ali, and M. A. Mohammed, "Vision-based approaches for automatic food recognition and dietary assessment: a survey," *IEEE Access*, vol. 7, pp. 35370–35381, 2019.

[3] F. Ahmed and C. Siwar, "Food intake and nutritional status among adults: sharing the Malaysian experience," *Pakistan Journal of Nutrition*, vol. 12, no. 11, pp. 1008–1012, 2013.

[4] NC Institute, "Obesity and cancer risk," 2017, https://www.cancer.gov/about-cancer/causes-prevention/risk/obesity/obesity-fact-sheet.

[5] E. J. Gallagher and D. LeRoith, "Obesity and diabetes: the increased risk of cancer and cancer-related mortality," *Physiological Reviews*, vol. 95, no. 3, pp. 727–748, 2015.

[6] World Health Organization, "Obesity study," 2011, http://www.who.int/mediacentre/factsheets/fs311/en/index.html.

[7] World Health Organization, "World Health Statistics 2012," 2012, http://www.who.int/gho/publications/world_health_statistics/2012/en/index.html.

[8] P. Pouladzadeh, P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi, "Food calorie measurement using deep learning neural network," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–6, Taipei, Taiwan, 2016.

[9] World Health Organization, "Obesity and overweight," 2020, https://www.who.int/news-room/fact-sheets/detail/obesityand-overweight.

[10] S. A. Ayon, C. Z. Mashrafi, A. B. Yousuf, F. Hossain, and M. I. Hossain, "FoodieCal: a convolutional neural network based food detection and calorie estimation system," in *2021 National Computing Colleges Conference (NCCC)*, pp. 1–6, Taif, Saudi Arabia, 2021.

[11] V. H. Reddy, S. Kumari, V. Muralidharan, K. Gigoo, and B. S. Thakare, "Food recognition and calorie measurement using image processing and convolutional neural network," *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 2019, pp. 109–115, Bangalore, India, 2019.

[12] S. Turmchokkasam and K. Chamnongthai, "The design and implementation of an ingredient-based food calorie estimation system using nutrition knowledge and fusion of brightness and heat information," *IEEE Access*, vol. 6, pp. 46863–46876, 2018.

[13] Z. Shen, A. Shehzad, S. Chen, H. Sun, and J. Liu, "Machine learning based approach on food recognition and nutrition estimation," *Procedia Computer Science*, vol. 174, pp. 448–453, 2020.

[14] J. Teng, D. Zhang, D. J. Lee, and Y. Chou, "Recognition of Chinese food using convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 9, pp. 11155–11172, 2019.

[15] P. Ma, C. P. Lau, N. Yu, A. Li, and J. Sheng, "Application of deep learning for image-based Chinese market food nutrients estimation," *Food Chemistry*, vol. 373, article 130994, Part B, 2022.

[16] P. Kadam, S. Pandya, S. Phansalkar et al., "FVEstimator: a novel food volume estimator wellness model for calorie measurement and healthy living," *Measurement*, vol. 198, article 111294, 2022.

[17] S. Podder, S. Bhattacharjee, A. Roy, and Department of Electronics, West Bengal State University, Barasat, Kolkata, India 700126, "An efficient method of detection of COVID-19 using Mask R-CNN on chest X-ray images," *AIMS Biophys*, vol. 8, no. 3, pp. 281–290, 2021.

[18] C. J. Burke, P. D. Aleo, Y. C. Chen et al., "Deblending and classifying astronomical sources with Mask R-CNN deep learning," *Monthly Notices of the Royal Astronomical Society*, vol. 490, no. 3, pp. 3952–3965, 2019.

[19] E. Karatoprak and S. Seker, "An improved empirical mode decomposition method using variable window median filter for early fault detection in electric motors," *Mathematical Problems in Engineering*, vol. 2019, Article ID 8015295, 9 pages, 2019.

[20] J. Zhang, X. Song, J. Feng, and J. Fei, "X-ray image recognition based on improved mask R-CNN algorithm," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6544325, 14 pages, 2021.

[21] "P. by Blogger, Indian food calorie chart," http://indianfoodrecipeswithpictures.blogspot.com/.