Bhutanese food recognition from images

Sonam Wangmo
Department of Electrical and Computer Engineering
Naresuan University
Phitsanalok, Thailand
sonamw65@nu.c.th

Panomkhawn Riyamongkol
Department of Electrical and Computer Engineering
Naresuan University
Phitsanalok, Thailand
panomkhawnr@nu.ac.th

Rattapoom Waranusast
Department of Electrical and Computer Engineering
Naresuan University
Phitsanalok, Thailand
rattapoomw@nu.ac.th

2024 23rd International Symposium on Communications and Information Technologies (ISCIT) | 979-8-3503-5395-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISCIT63075.2024.10793693

Abstract— Despite the popularity of using deep learning techniques in food recognition, object detection, and classification in the context of Bhutanese food remains unexplored. This study presents the recognition of Bhutanese food from images using yolov8 models. The paper introduces a new Bhutanese food dataset called BHT-Food22, consisting of 22 Bhutanese dishes with 100 images in each category. The food images were collected from a few volunteers' daily food consumption and the internet. The dataset was randomly divided into 80% training and 20% validation sets. Data augmentation techniques such as scaling, flipping, translation, rotation, mosaic, and HSV color-space adjustment were applied during the training process. The dataset was trained on YOLOv8s-seg model, which uses CSPDarknet53 for feature extraction and object detection. The model was able to achieve an average mAP@50 of 0.91 for food detection and mAP@50 of 0.908 for food segmentation in the validation set.

Keywords— Bhutanese Food, Food Recognition, Deep Learning, YOLO, BHT-Food22

I. INTRODUCTION

Bhutan is a small landlocked country encircled by two giant countries, China in the north and India in the south. With a small geographical area and population, the security of the sovereign of the country has always been the priority of the nation. Apart from providing nutrients to the human body, food embodies deeper cultural meaning and personal history [1]. It also serves as a symbol of national identity and heritage, playing a crucial role in preserving the country's unique culture. Despite the popularity of using deep learning techniques in the past decade to recognize food, the object detection and recognition on Bhutanese food remains unexplored. Moreover, there is no available food dataset that contains Bhutanese cuisines. Some of the popular food datasets like UECFOOD-100 [2], and UECFOOD-256 [3] included Japanese cuisines, ETHZ Food-101 [4], Food-101N [5] included Western food, Vireo Food-172 [6], ChineseFoodNet [7] included Chinese dishes and ISIA Food-500 [8] covered food from various countries and regions including Western and Eastern cuisine.

Compared to general image recognition tasks, food recognition tasks are inherently challenging as there is high variation in intra-class and low variation in inter-class food types. This is caused due to different cooking methods which led to diversity in shape, color, and texture [9], [10]. Similar to Indian food platter [10], Bhutanese food does not have distinct boundaries between dishes which makes recognition of food more challenging. Delving into food details can help in creating awareness on nutrition and dietary health. By using photos of Bhutanese dishes as input and generating dish names as outputs, the system not only helps people learn about Bhutanese cuisines but also indirectly promotes the country's cultural heritage.

In this paper, we used YOLOv8 to train a model for detecting, segmenting, and recognizing Bhutanese food. We created a new dataset containing 22 types of local and traditional Bhutanese dishes.

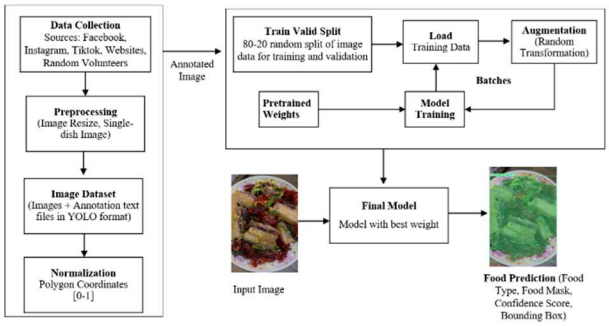

Fig. 1. System Workflow

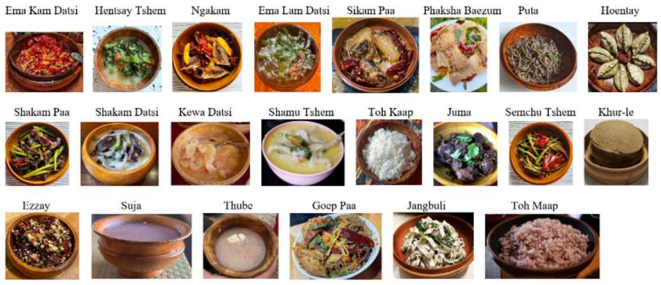

Fig. 2. Bhutanese Dishes in BHT-Food22

II. RELATED WORK

For food recognition, many researchers had used traditional approaches like kNN (k-Nearest Neighbor) and SVM (support vector machine). Phetphoung, Kittimeteeworakul and Waranusast [11] developed a system to classify types of sushi by extracting color and shapes features of sushi from camera, and used kNN classifier to recognize and detect different types of sushi. Similarly, dish images in [12] were recognized by extracting scale-invariant feature transform (SIFT), color and circle edge features, and

used SVM classifier to classify the food. However, Sombutkaew and Chitsobhuk used Mask R-CNN pre-trained on CoCo dataset to retrain Thai food dataset consisting of 11 food categories, where the model could correctly classify the food with an accuracy of 0.984. The segmented image of Thai food was further processed to predict the calorie of segmented food [13]. Similarly, [14] compared two-stage Mask RCNN and one-stage YOLOv8 for instance segmentation under various orchard conditions. The result showed that YOLOv8 performed better than Mask RCNN in both segmentation and inference times.

In-depth investigation on food detection, classification and analysis was done in [9]. The authors states that CNN is used extensively in food detection as it provided better accuracy compared to other models. Nevertheless, the authors [15] observed that hybrid vision transformer outperforms the most sophisticated ensemble CNN architecture in categorizing the food. They used 13 food categories with 400 images of Indian food. The recommended model scored an accuracy of 94.63%, specificity of 95.23%, sensitivity of 84.42% and kappa coefficient of 0.93.

Likewise, the [16] created dataset from Food-11, FooDD, Food100 datasets and web archives. They designed a convolutional neural network model from scratch and compared its performance with the fine-tuned pretrained models ALexnet and Caffenet. The result showed that finetuned models from transfer learning provided better accuracy than the designed structure.

An Indian-Food10 dataset [17] was created from 10 traditional Indian food items with more than 12,000 images. The researchers pre-trained the data on yolov4 object detection model and used the pre-trained weight on training the image dataset where they could achieve MAP score of 91.76% and F1 score of 0.90.

Yolov8 model has a slightly higher accuracy score than yolov7 in detecting the object [18]. The experiment conducted in [19] showed that yolov8 image detection model has higher accuracy than that of yolov8 image classification model in identifying Chinese herbal medicine slices. The paper used yolov8n model which is the smallest model among the yolov8 series due to its size of 5.94MB and low hardware requirements. The model generated an mAP@0.5 of 95.8%.

TABLE I. AUGMENTATION PARAMETER CONFIGURATION

Hyperparameters	Settings
hsv_h	0.015
hsv_s	0.7
hsv_v	0.4
degrees	45
translate	0.1
scale	0.5
fliplr	0.5
mosaic	1.0

III. METHOD

The system overflow of the entire study is depicted in Fig. 1, encompassing various stages such as data collection, preprocessing, data annotation and normalization, splitting of data into training and validation sets and finally training the model with data that are augmented while loading in batches. The dataset was experimented with YOLOv8s-seg and it was used as a basic framework to train the model. By integrating semantic segmentation into the detected input image, the YOLOv8s-seg model enhances the functionalities initially introduced by the YOLOv8 object detection model.

A. Data Collection

Bhutanese cuisine encompasses diverse dishes from all four regions of Bhutan. However, for this study, a new dataset was created with 19 traditional Bhutanese dishes and to increase the dataset, some of the popular local dishes from the central part of Bhutan were incorporated. The dataset was selected based on the popularity and availability. Finally, 22 dishes were selected to continue with data collection.

Most of the food images were accumulated from social networking sites like Facebook, Instagram, and TikTok. The search for the foods was done using hashtags and search terms like #bhutanesefood, #bumtapfood, and Bhutanese popular dishes. Few volunteers contributed their daily food consumption photos using mobile cameras where only Bhutanese cuisines consumed were saved and others were discarded manually.

The collected data contained a mixture of images showing single-dish and multi-dish. During data collection, there were no size limits which increased the data memory size. To decrease the memory size and to load the images faster during training of the model, the images were resized. Each multi-dish images were made into a single-dish image by covering the unwanted dish with black pixels. Finally, the BHT-Food22 dataset (Fig. 2) consists of 22 categories of Bhutanese food with 100 images in each category.

B. Data Annotation

The preprocessed data is labeled using the Roboflow platform. Roboflow is a computer vision platform that allows users to upload custom datasets, draw annotations, perform augmentation, and train and deploy computer vision models. The final image dataset consists of images with annotated text files in yolo format.

C. Data Splitting

The dataset includes 2200 images and is divided randomly into two sets: 80% as a training set and 20% as a validation set. The training set includes 1760 images and the validation set includes 440 images.

D. Data Augmentation

The data augmentations are applied directly to the dataset during the training process eliminating the need for creating additional images. Some of the techniques applied include adjustments to hue, saturation, and brightness to introduce variations in color distribution, contrast and lighting conditions. Image rotation was configured by specifying degrees value, while translation augmentation shifts or moves the object within the image. Scaling resizes the input image, and fliplr flips the image from left to right. The mosaic augmentation involves combining four input images into

single mosaic images, thereby creating diverse samples. This ensures that the model sees slightly different variation of images at every epoch. Table I gives the augmentation hyperparameters configured for the training of the model.

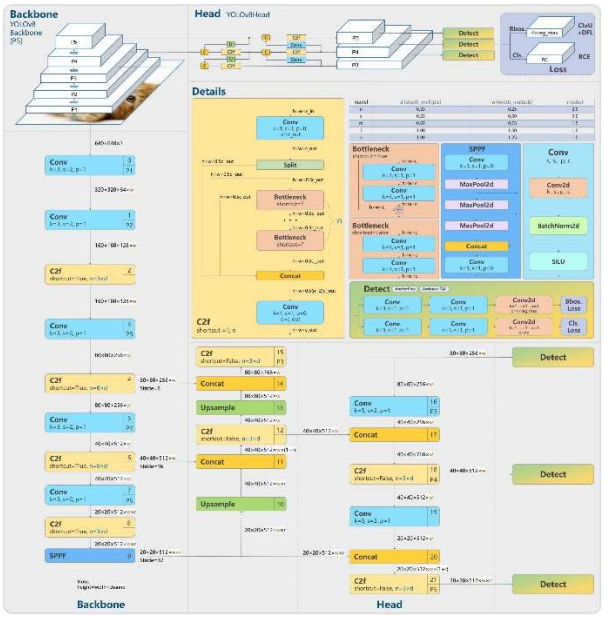

Fig. 3. YOLOv8 architecture for object detection and segmentation from RangeKing's GitHub repository

E. YOLOv8 Architecture

YOLOv8 is a cutting-edge, state-of-the-art (SOTA) model, launched on January 2023 by Glenn Jocher at Ultralytics. YOLOv8 is a one-stage model that supports pose estimation, detection, tracking, segmentation, and classification. YOLOv8 consists of versions of architecture; YOLOv8n is the lightest model and YOLOv8x is the heaviest model with 37.3 and 53.9 mAP scores respectively.

The core architecture of the YOLOv8-seg model (Fig. 3) is CSPDarknet53. It comprises 53 layers of a convolutional neural network that are designed for feature extraction and object detection tasks. It is followed by a novel C2f module which is a replacement of the C3 module of traditional YOLO. After the C2f module, two segmentation heads are trained to generate semantic segmentation masks for the input image. The model shares detection heads similar to those of YOLOv8, consisting of five detection modules and a prediction layer. YOLOv8 employs a decoupled head where the head no longer performs classification and regression together. It separates the task of object detection and classification, improving the model's performance in both accuracy and processing speed. The SPPF layer and the following convolution layers process features at various scales, while the Upsample layers enhance the resolution of the feature maps. To improve detection accuracy, the C2f module combines contextual data with high-level features.

Finally, the detection module utilizes a sequence of convolutional and linear layers to map the high-dimensional features to the output bounding boxes and object classes [20].

F. Performance Evaluation

Precision, Recall, mAP, and F1-Score were among the metrics used to assess the model's performance. Each metric is defined by its respective equation (1)-(4).

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$mAP = \left(\frac{1}{K}\right)\sum_{i=0}^{k}(AP)i \tag{3}$$

$$F_{1-score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{4}$$

Here TP, FP denotes true and false positive, and FN refer to false negative of object instances. 'K' is the total number of classes, and (AP)i refers to the average precision for each category.

Table II displays the confusion matrix used in classification problems to describe the errors in the model. It has four different quadrants: TP, FP, FN, and TN. TP indicates the number of positive observations correctly predicted. True Negative (TN) is negative observation correctly predicted as negative samples. FP is negative observation detected as positive. FN is the number of positive samples detected as negative.

TABLE II. CONFUSION MATRIX

	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

TABLE III. EXPERIMENTAL ENVIRONMENT

No.	**Category**	**Settings**
1	Operating System	Windows 11 Pro
2	Development environment	Python-3.11.8 torch-2.2.2+cu118, Ultralytics YOLOv8.1.34, CUDA 12.2
3	Graphics card (GPU)	NVIDIA GeForce RTX 4070 12282MiB (12G/ASUS ROG G22CH)

(a) (b)

(c)

Fig. 4. Dish image for inter-class similarities in (a) Ema Lam Datsi and Semchu Tshem (b) Ema Lam Datsi and Ema Kam Datsi and intra-class variation in (c) variation of Goep Paa

TABLE IV. PERFORMANCE EVALUATION FOR EACH FOOD CATEGORY

Food Images	Food Category	Precision (B)	Recall (B)	mAP@0.5 (B)	Precision (M)	Recall (M)	mAP@0.5 (M)
	all	0.882	0.856	0.91	0.879	0.854	0.908
	Ema Kam Datsi	0.942	0.857	0.942	0.942	0.857	0.942
	Ezzay	0.981	0.95	0.988	0.98	0.95	0.988
	Ema Lam Datsi	0.639	0.8	0.825	0.639	0.8	0.825
	Kewa Datsi	0.834	0.8	0.858	0.834	0.8	0.858
	Nga Kam	0.762	0.909	0.813	0.762	0.909	0.813
	Toh Maap	0.979	0.905	0.955	0.979	0.905	0.955
	Shakam Paa	0.941	0.802	0.921	0.941	0.802	0.921
	Sikam Paa	0.723	0.95	0.911	0.723	0.95	0.911
	Suja	0.85	0.95	0.933	0.85	0.95	0.933
	Hentsay Tshem	0.793	0.9	0.95	0.792	0.9	0.95
	Semchu Tshem	0.847	0.754	0.814	0.847	0.754	0.814
	Jangbuli	0.907	0.932	0.934	0.907	0.932	0.934
	Hoentay	0.869	0.57	0.66	0.869	0.57	0.663
	Goep Paa	0.931	0.672	0.926	0.931	0.672	0.926
	Thube	0.906	0.88	0.962	0.906	0.881	0.962
	Khur-le	0.924	0.95	0.986	0.923	0.95	0.986
	Juma	0.954	0.818	0.923	0.901	0.773	0.889
	Phaksha Baezum	0.945	0.853	0.941	0.945	0.853	0.941

	Shakam Datsi	0.952	1	0.995	0.952	1	0.995
	Shamu Tshem	0.844	0.814	0.898	0.844	0.814	0.898
	Toh Kaap	0.919	0.947	0.975	0.919	0.947	0.975
	Puta	0.962	0.818	0.906	0.962	0.818	0.906

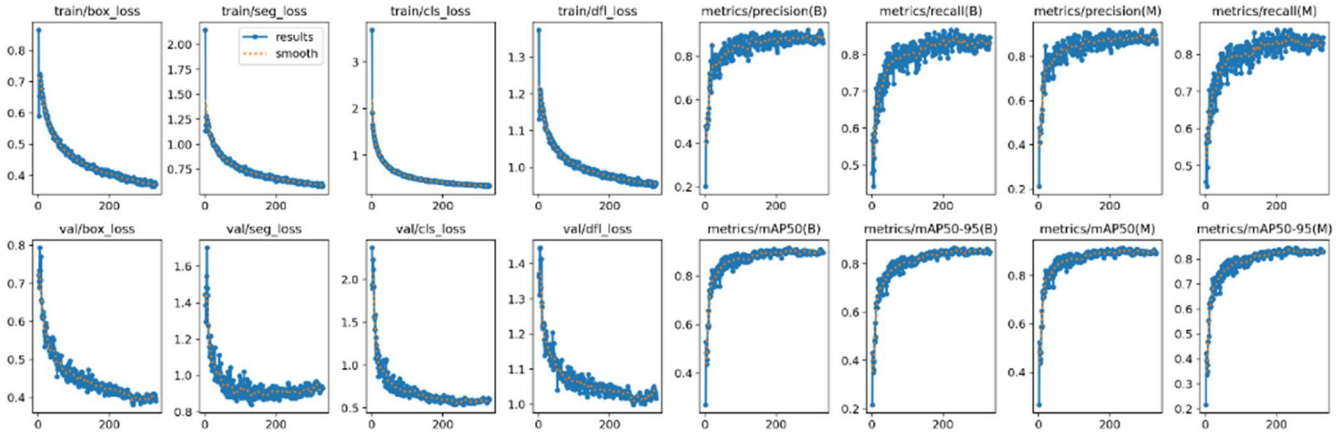

Fig. 5. The result for detection, classification and segmentation of food from YOLOv8 model on training and validation set.

IV. EXPERIMENTAL RESULT ANALYSIS

The model was trained with a batch-size of 16, and an image-size of 640x640 pixels for 1000 epochs. The model underwent training for 331 epochs, halting when further improvements in learning were not observed. To prevent overfitting issues during training, early stopping was implemented with a patience parameter of 100. Table III outlines the computer configuration utilized for this study's training process. The model evaluates its performance on the validation dataset during the training process after each epoch. The precision, recall, and mAP values for each class for validation set are shown in Table IV.

Fig. 5 illustrates the training results, depicting both training loss and validation loss. To assess the model performance, four different loss values were evaluated at each epoch: box loss (box_loss), distribution focal loss (dfl_loss), segmentation loss (seg_loss) and classification loss (cls_loss). The box loss measures how well the predicted bounding boxes match the ground truth bounding boxes. The dfl loss helps model estimate the offsets to improve the accuracy of localizing objects within an image. Meanwhile, classification loss measures the correctness of model in classifying the detected object. The segmentation loss indicates how well the model differentiates between objects pixel and background pixels. Similarly, the last four graphs are the mAP50 and mAP50-95 values for bounding boxes and masks, respectively.

Fig.6 illustrates the confusion matrix for each class where the model incorrectly predicted 0.10% of Hentsay Tshem, Kewa Datsi, and Shamu Datsi as Ema Lam Datsi. Similarly, 0.10% of Shakam Paa was misclassified as Ngakam, and 0.20% of Goep Paa was misclassified as Sikam Paa. These mistakes are mainly due to the similarities between different food classes and variations within the same class. For instance, Fig. 4 (a, b) demonstrates the inter-class similarities and Fig. 4c displays intra-class variations. Fig4a shows Ema Lam Datsi and Semchu Tshem. Fig4b shows Ema Lam Datsi and Ema Kam Datsi and Fig 4c shows variation of Goep Paa dishes. The background of the food images also played a role in these wrong predictions, as some dishes were mistaken for the background and vice versa. The Hoentay class had the highest rate of being wrongly identified as background, at 0.40%. Fig. 7 provides examples of these misclassifications, showing instances where Hoentay was mistaken for the background (Fig7a) and the background was mistaken for Hoentay (Fig7b). With the variety of folding techniques, Hoentay (dumplings) exhibits diverse and irregular shapes unlike objects with more consistent shapes. In addition, Hoentay are often stacked or clustered on plates, leading to occlusion. This means parts of some Hoentay are hidden by others, which can make it difficult for the model to differentiate individual Hoentay and predict correct bounding boxes or segmentations.

The result shows that the model performance for the class Hoentay was not good, however, the model achieved an average mAP@50 of 0.91 for food detection and mAP@50 of

0.908 for food segmentation depicting a slightly better score for object detection than object segmentation.

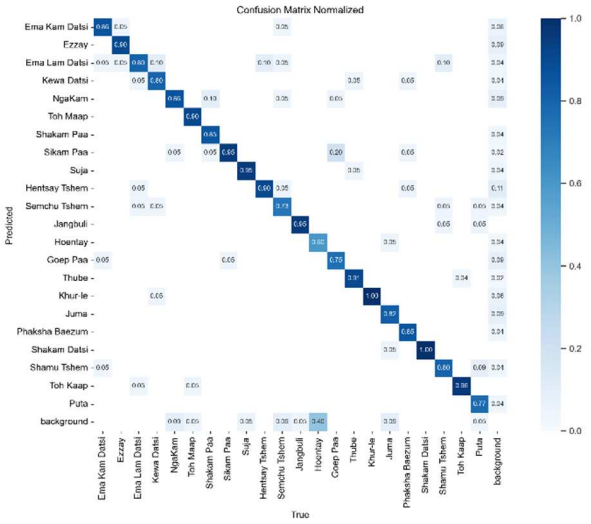

Fig. 6. Confusion Matrix for 22 Classes

(a) (b)

Fig. 7. Wrong prediction of Hoentay as background in (a) and background predicted as Hoentay in (b)

V. Conclusion

This paper introduced a new dataset for Bhutanese dishes of 22 types of popular food in Bhutan. BHT-Food22 consists of 2200 food images. The YOLOv8s-seg model was used in this study to detect objects and perform semantic segmentation of the detected object in the image. Based on the experiment, the model could perform well on the Bhutanese food images; however, the model can be further improved by incorporating more samples of data and employing advance algorithms to improve its accuracy and performance. Additionally, more types of dishes from other regions of Bhutan can be incorporated in future studies. This study can be extended further by developing an application on mobile devices to recognize Bhutanese food. With a larger dataset and improved performance, it can offer other services like determining the food portion to monitor calorie content and estimate the shape and size of the food.

Acknowledgment

S., P. and R. would like to extend their sincere gratitude to Naresuan University for providing resources and support. S, a scholarship recipient, wishes to express her thankfulness to His Majesty the King of Bhutan and Naresuan University for granting her a master's degree scholarship.

References

[1] L. He, Z. Cai, D. Ouyang and H. Bai, "Food Recognition Model Based on Deep Learning and Attention Mechanism," *2022 8th International Conference on Big Data Computing and Communications (BigCom)*, Xiamen, China, 2022, pp. 331-341, doi: 10.1109/BigCom57025.2022.00048.

[2] Y. Matsuda and K. Yanai, "Multiple-food recognition considering co-occurrence employing manifold ranking," *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Tsukuba, Japan, 2012, pp. 2017-2020.

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

[4] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in European Conference on Computer Vision. Springer, 2014, pp. 446–461.

[5] K. -H. Lee, X. He, L. Zhang and L. Yang, "CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5447-5456, doi: 10.1109/CVPR.2018.00571.

[6] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 32–41.

[7] X. Chen, H. Zhou, Y. Zhu, and L. Diao, "Chinesefoodnet: A large-scale image dataset for chinese food recognition," 2017, arXiv:1705.02743. [Online]. Available: https://arxiv.org/abs/1705.02743

[8] W. Min et al., "ISIA food-500: A dataset for large-scale food recognition via stacked global-local attention network," in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 393–401

[9] A. Banerjee, P. Bansal, and K. T. Thomas, "Food Detection and Recognition Using Deep Learning – A Review," in 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 16-17 Dec. 2022 2022, pp. 1221-1225, doi: 10.1109/ICAC3N56670.2022.10074297.

[10] D. Pandey et al., "Object Detection in Indian Food Platters using Transfer Learning with YOLOv4," arXiv (Cornell University), May 2022, doi: 10.1109/icdew55742.2022.00021.

[11] W. Phetphoung, N. Kittimeteeworakul, and R. Waranusast, ''Automatic sushi classification from images using color histograms and shape properties,'' in Proc. 3rd ICT Int. Student Project Conf. (ICT-ISPC), Mar. 2014, pp. 83–86, doi: 10.1109/ICTISPC.2014.692322.

[12] K. Kitamura, C. De Silva, T. Yamasaki, K. Aizawa, "Image processing based approach to food balance analysis for personal food logging," In Proc. IEEE int. Corif. on Multimedia and Expo, 20 I 0, pp. 625 -630.

[13] R. Sombutkaew and O. Chitsobhuk, "Image-based Thai Food Recognition and Calorie Estimation using Machine Learning Techniques," 2023 20th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Nakhon Phanom, Thailand, 2023, pp. 1-4, doi: 10.1109/ECTI-CON58255.2023.10153183.

[14] R. Sapkota, D. Ahmed, and M. Karkee, "Comparing YOLOv8 and Mask RCNN for object segmentation in complex orchard environments," arXiv.org, Dec. 27, 2023. https://arxiv.org/abs/2312.07935.

[15] R. Nijhawan, A. Batra, O. Loyola-Gonz´alez, M. Kumar, and D. K. Jain, "Food Classification of Indian Cuisines Using Handcrafted

Features and Vision Transformer Network," SSRN Electronic Journal, 2022, doi: 10.2139/ssrn.4014907.

[16] Gozde Ozsert Yıgıt and Buse Melis Özyildirim, "Comparison of convolutional neural network models for food image classification," Jul. 2017, doi:10.1109/inista.2017.8001184.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934 [cs, eess], Apr. 2020, Available: https://arxiv.org/abs/2004.10934v1

[18] Zineb Haimer, Khalid Mateur, Y. Farhan, and Abdessalam Aït Madi, "Pothole Detection: A Performance Comparison Between YOLOv7 and YOLOv8," Oct. 2023, doi: 10.1109/icoa58279.2023.10308849.

[19] Y. Su, B. Cheng, and Y. Cai, "Detection and Recognition of Traditional Chinese Medicine Slice Based on YOLOv8," Jul. 2023, doi: 10.1109/iceict57916.2023.10245026.

[20] G. Jocher, A. Chaurasia, and J. Qiu, "*YOLOv8 by Ultralytics*," GitHub, Jan. 01, 2023. https://github.com/ultralytics/ultralytics