Grenze International Journal of Engineering and Technology, June Issue

NutriScan: An Estimation of Nutrition and Volume of Food Item through Monocular Image Analysis

Shreya N Palavalli[1], Vibha Jamadagni[2], Shrithik Krupakara Reddy[3], Arshya Khurana[4] and Chitra G M[5]
[1-5]PES University/Dept. of Computer Science, Bengaluru, India,
Email: {shreyapalavalli, vibhajamadagni, shrithikreddy11122002, arshyakhurana}@gmail.com, chitragm@pes.edu

***Abstract*— With the growing concern over diet-related health challenges, there is an urgent need for innovative tools to simplify and enhance the way we monitor and understand food consumption. This study introduces an innovative image-based system to estimate food volume and nutritional content from single monocular images, leveraging advanced computer vision and deep learning techniques for accurate food classification, volume estimation, and nutritional profiling. The approach aims to offer a robust and adaptable solution for dietary assessment, personalized nutrition, and efficient health tracking. By integrating food classification using a pre-trained Inception V3 model with precise volume estimation through depth analysis and segmentation, the system achieves robust and dependable outcomes. A custom segmentation model isolates the food region, while calibrated depth information ensures reliable volume measurements. The proposed framework holds potential for applications in dietary monitoring and health management by providing an efficient, image-based solution for analyzing food consumption.**

***Index Terms*—Computer Vision, Deep Learning, Nutritional Analysis, Food Volume Estimation, Dietary Monitoring.**

I. INTRODUCTION

The global increase in diet-related health issues, such as obesity, diabetes, and cardiovascular disease to name a few, have highlighted the need to monitor food intake. Traditional methods of assessing food portions, often relying on manual measurements or specialized equipment, are not only time-consuming, but are also prone to errors. These methods are impractical for everyday use, particularly for people such as health conscious consumers, patients with specific dietary needs, or professionals in the food industry. As such, there is an urgent demand for more efficient, scalable solutions that can provide reliable nutritional information in real time without the need for complex tools or expertise.

This growing need for reliable and accessible food analysis has led to the exploration of modern technologies, particularly in the fields of computer vision and deep learning. These technologies offer the potential to revolutionize the way we assess food, moving beyond the limitations of traditional methods. With the ability to analyze food images and infer both volume and nutritional information automatically, computer vision systems promise to make dietary monitoring more accurate and scalable. However, despite the growing interest in these technologies, most existing systems still face challenges in providing real-time, reliable, and comprehensive food analysis in everyday contexts.

Grenze ID: 01.GIJET.11.2.557_1

This paper introduces a framework that combines advanced image analysis and deep learning techniques to address these challenges. The proposed solution automates the estimation of both the physical volume and the nutritional profile of food items, using only a simple image as input. At the heart of this framework is a pre-trained Inception V3 model, which is utilized to classify food items and retrieve their corresponding nutritional data. In parallel, a depth estimation model, Depth Anything V2 along with a segmentation model, yolo11, analyzes the image to derive a 3D structure of the food. This 3D model is then used to calculate the volume of food, which is crucial for a reliable estimation of the portion size.

With its potential to integrate into mobile applications, healthcare devices, and even industrial settings, this paper represents a significant step toward making food analysis faster, more reliable, and widely accessible. Using advanced technology to solve a real-world problem, it empowers people and professionals alike to make more informed decisions about their nutrition.

II. Related Work

Advances in dietary assessment have been fueled by the need to improve accuracy in portion size estimation and nutrient analysis, addressing the challenges of traditional self-reported methods. Gao et al. [1] developed a comprehensive system for nutrient extraction based on image recognition and entity extraction. Their system combined the FoodData Central (FDC) nutrient database with machine learning for food identification. By evaluating three architectures—Inception-v3, Inception-v4, and MobileNetV2—they found MobileNetV2 achieved the highest accuracy (96.71%) with a lightweight computational footprint. Additionally, a custom Named Entity Recognition (NER) model using SpaCy achieved an F-score of 88.15%. However, the system faced challenges related to limited training data diversity and missing nutrient information in the FDC database.

Kim et al. [2] proposed a method for food classification and meal intake estimation using pre- and post-meal images. They employed Mask R-CNN with ResNet-50 for pixel-level detection, enabling nutrient analysis through food type classification and intake volume estimation. Despite achieving high accuracy in classification (97.57%) and detection (93.6%), the method relied on simplified 3D shape modeling for volume estimation, which limited accuracy for irregularly shaped foods and mixed meals.

Shao et al. [3] introduced a monocular image-based approach for food portion size estimation, leveraging a conditional GAN (cGAN) for 3D voxel reconstruction, a deep regression model for energy density estimation, and a refinement module for volume predictions. While demonstrating competitive accuracy on the Nutrition5k dataset, the reliance on 3D voxelization introduced potential error propagation, especially with inaccurate depth maps or segmentation errors.

Rahman et al. [4] explored various image-based dietary assessment systems, including the Technology Assisted Dietary Assessment (TADA) project, which used fiducial markers for automated volume estimation. Shang et al. [10,11] advanced the field with structured light-based approaches for depth mapping and volume estimation but required external hardware like lasers. Single-image methods, such as those using predefined shapes, showed promise but struggled with irregularly shaped foods and unconventional conditions.

Dehais et al. [5] discussed multi-view approaches for 3D food volume reconstruction, emphasizing advancements like DietCam, which used sparse 3D point clouds generated from multiple images. Despite improved accuracy, these methods required significant user effort and hardware, making them less practical for real-world applications.

Graikos et al. [6] proposed a depth prediction and segmentation system for food volume estimation using a single image. Their self-supervised depth estimation network, trained on egocentric food-handling videos, achieved reduced errors after fine-tuning on a custom dataset. The system demonstrated the capability to estimate volumes for multiple food items simultaneously, but challenges remained with occluded foods, unconventional capturing angles, and scaling inaccuracies.

These studies highlight the progress in automated dietary assessment, with key contributions ranging from improved image recognition techniques to novel volume estimation methods. However, limitations such as reliance on predefined shape models, external hardware, and inconsistent performance under real-world conditions underscore the need for further advancements.

III. Dataset

A. Food Nutrition Dataset

The Food Nutrition Dataset comprises an extensive collection of detailed nutritional information for eight specific food classes: Biryani, Chapati, Dosa, Idli, Noodles, Thatte Idli, Upma, and Ven Pongal. This dataset

uniquely targets the South Indian demographic, making it a pioneering and specialized resource in the field. By focusing on these traditional foods, it opens opportunities for researchers and culinary enthusiasts to delve into a diverse array of cuisines and conduct comprehensive studies on dietary patterns. For each food item, the dataset meticulously presents calorie content along with macro-nutrient values, including proteins, fats, and carbohydrates, quantified per 100 grams along with their densities. This well-structured information facilitates precise calculations of total nutrient intake based on the identified food items. By concentrating on these eight classes, the dataset ensures accurate and reliable nutrient analysis, which is indispensable for comprehending dietary consumption within this specific cultural context.

B. Food Image Dataset

The Food Image Dataset is a self-curated compilation of 716 high-quality images that represent the same eight distinct food classes, specifically showcasing South Indian cuisine. These images have been enhanced through meticulous preprocessing techniques, including data augmentation strategies such as noise addition and horizontal flipping, which equip the model to recognize foods under varied conditions. The dataset has been rigorously annotated using Roboflow, a sophisticated image annotation tool that streamlines the labeling process. Following annotation, the images were systematically divided into training, validation, and test sets, facilitating efficient model training and robust performance evaluation. By emphasizing South Indian culinary traditions, this dataset not only enriches the existing body of knowledge but also fosters exploration and investigation into other regional cuisines, paving the way for a deeper understanding of their nutritional attributes and cultural relevance.

Figure 1. Food Image Dataset

IV. METHODOLOGY

The framework involves several stages: image classification, segmentation, depth estimation, 3D reconstruction, and volume calculation.

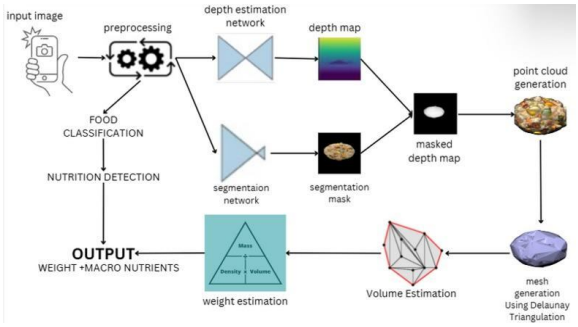

Figure 2. Architecture Diagram

Figure 3. Food Classification

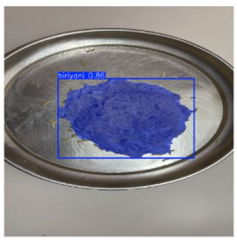

Figure 4. Food Segmentation

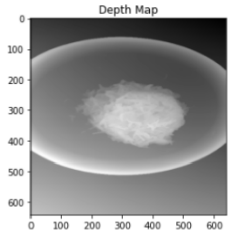

Figure 5. Depth Map Generation

A. Food Classification

The first step in the process is to recognize food items from images as shown in Figure 3, where both Inception V3 and YOLO11 were tested for image classification. Inception V3 is a convolutional neural network pretrained on the ImageNet dataset, which contains millions of labelled images, making it highly effective for image classification tasks due to its efficient feature extraction and accurate categorization capabilities. YOLO11, on the other hand, is primarily designed for real-time object detection and is pre-trained on the COCO dataset, which includes a diverse range of object categories. Both models were fine-tuned using our custom food image dataset, and it was observed that Inception V3 achieved significantly better classification results, with an accuracy of 96.88%, compared to 84.615% obtained by YOLO11. For Inception V3, the model was fine-tuned using Adam optimizer with a learning rate of 0.001, which produced the best results. The input images were resized to a target size of (228, 228) and a batch size of 32 was used. The training was done for 20 epochs as the accuracy improvements stabilized, indicating that optimal performance had been reached.

B. Segmentation of food item from image

The next step after food recognition is segmenting the food item from the image as shown in Figure 4, for which YOLO11 is utilized. YOLO11, the latest and most advanced model for object detection and segmentation, is renowned for its real-time performance and high accuracy. It excels in identifying and segmenting objects within complex scenes, making it an ideal choice for this task. To tailor the model for the specific requirements of food segmentation, YOLO11 was fine-tuned on our custom food image dataset. This fine-tuning ensured that the model accurately identified and segmented food items. Segmentation isolates the target food items, preparing them for further depth and volume estimation.

C. Depth map generation for the 2D image

In the depth generation phase, the Depth Anything V2 model is used for monocular depth estimation as seen in Figure 5. Depth Anything V2 surpasses previous models like MiDaS and Depth Anything V1 by delivering more accurate and efficient depth estimations. This state-of-the-art model predicts the relative depth of each point in an image. This process involves estimating a depth map D from an input image I, where each pixel (i, j) in D represents the relative distance of the corresponding point in I from the camera. The monocular depth estimation is formulated as finding a function f such that:

$$D = f(I)$$

Depth Anything V2 enhances this estimation by utilizing a convolutional neural network (CNN) architecture trained on a vast dataset comprising 595,000 synthetic labeled images and over 62 million real unlabeled images. This extensive training enables the model to learn complex features and relationships within images, leading to finer and more robust depth predictions. The model's architecture includes an encoder-decoder structure, where the encoder extracts hierarchical features from the input image, and the decoder reconstructs the depth map. The loss function L used during training combines photometric loss Lp, which measures the difference between the input image and a synthesized image from the estimated depth, and smoothness loss Ls, which enforces smoothness on the depth map:

$$L = Lp + \lambda Ls$$

Here, λ is a weighting factor that balances the two loss components.

In this project, Depth Anything V2 was fine-tuned on our custom food image dataset to ensure precise depth estimation tailored to the specific characteristics of food items. The resulting depth maps provide critical 3D spatial information as shown in Figure 5, which facilitates accurate modeling of the physical structures of the food items.

D. 3D Reconstruction of the food item

The segmented depth data is converted into a point cloud using the Point Cloud Library (PCL) as shown in Figure 6, an open-source framework designed for efficient 2D/3D image and point-cloud processing. This transformation involves mapping the 2D depth data D(x, y) into 3D coordinates (X, Y, Z) in the camera coordinate system.

The transformation follows the perspective projection model:

$$\boldsymbol{X} = \frac{(\boldsymbol{x} - \boldsymbol{c_x}) \cdot \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})}{\boldsymbol{f_x}},$$

$$Y = \frac{(y - c_y) \cdot D(x, y)}{f_y},$$

$$Z = D(x, y),$$

where:

- (x, y) are the pixel coordinates in the 2D depth map.
- D(x, y) is the depth value at pixel (x, y).
- (c_x, c_y) are the principal point offsets in the camera intrinsic matrix.
- (f_x, f_y) are the focal lengths along the X- and Y- axes, respectively.

This mapping generates a dense 3D representation of the food item, capturing intricate surface details critical for accurate volume estimation. The resulting point cloud is processed to remove noise and outliers using the Statistical Outlier Removal (SOR) filter, ensuring a high-fidelity representation of the object's geometry.

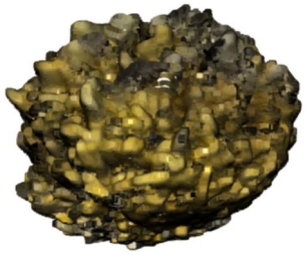

Figure 6. Point Cloud Generation

Figure 7. Mesh Generation

E. Mesh generation and Volume estimation of the food item

To estimate the volume of the food item, the processed point cloud is used to generate a 3D mesh as seen in Figure 7. This is accomplished by applying α-complex filtration to the point cloud, which retains simplices (points, edges, triangles, and tetrahedrons) based on a specified α-value. An α-value of 0.01 is used here. The α-complex technique efficiently divides the point cloud into a series of tetrahedrons, forming a detailed 3D representation of the food item's surface.

The volume of each tetrahedron in the mesh is calculated using the scalar triple product, which determines the signed volume of a tetrahedron given its four vertices. The formula for the volume of a tetrahedron with vertices a,b,c,d is:

$$V = \frac{1}{6}\left|(a - d) \cdot \left((b - d)\ x\ (c - d)\right)\right|$$

Here, (a-d)·((b-d) x (c-d)) represents the scalar triple product, and dividing by 6 accounts for the tetrahedron's geometric properties. The total volume of the food item is obtained by summing the volumes of all tetrahedrons filtered by the α-complex. The weight of the food item is then determined by multiplying this total volume by the food's density.

F. Nutritional profiling of the food item

A custom food nutrition dataset containing detailed nutritional profiles per 100 grams is used to provide the final nutritional breakdown. Based on the calculated weight of the food item, its nutritional content (macro-nutrients and calories) was estimated and presented to the user.

```
Nutritional values for 204.80 grams of Biriyani:
  Protein (g): 20.48
  Carbohydrates (g): 92.16
  Fats (g): 30.72
  Calories (kcal): 798.71
  Fiber (g): 6.14
  Sugar (g): 4.10
```

Figure 8. Nutritional Profiling

This systematic methodology integrates deep learning and computer vision techniques, providing an efficient and accurate way to estimate food volume and nutritional information from images.

V. RESULTS

To validate volume estimation, we compared predicted volumes—converted to weight—with physical measurements using a precision scale. Our model converts segmented depth data into a 3D point cloud using the Point Cloud Library (PCL). This is further processed using an α-complex filtration technique to generate a 3D mesh, which allows for accurate volume computation by summing the volumes of tetrahedral segments. The estimated volume is then multiplied by the food's density to determine its weight. This approach ensures high precision in food volume and weight estimation. The calculated weight was then compared to the actual weight measured on the scale to assess the accuracy of our approach. Additionally, we report the Mean Absolute Percentage Error (MAPE) values in our results table to quantify the deviation between predicted and actual weights.

TABLE 1. ESTIMATED WEIGHT, VOLUME AND MEAN ABSOLUTE PERCENTAGE ERROR

Food	Actual Weight(g)	Predicted weight(g)	Volume(cm^3)	MAPE
Biryani	203	214.9	202.73	5.86%
Noodles	292	373.27	339.34	27.83%
Upma	108	181	164.55	67.59%
Thatte Idli	61	89.12	148.53	46.1%
pongal	201	238.3	216.6	18.56%

Each individual component of the system was tested for its specific task. We used two models for food classification—InceptionV3 and YOLO11—and compared their accuracies. InceptionV3 achieved an accuracy of 96.88%, while YOLO11 achieved 84.615%. This comparison highlights the superior performance of InceptionV3 in classifying food items, likely due to its deeper architecture and efficient feature extraction. However, YOLO11 provides a faster, real-time classification capability, making it useful for applications requiring speed over absolute accuracy.

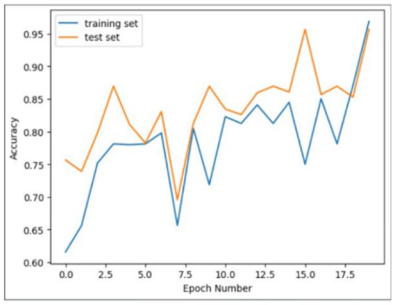

Figure 9. InceptionV3 accuracy graph

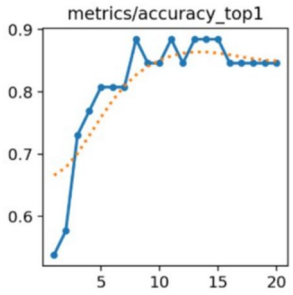

Figure 10. YOLOv11 accuracy graph

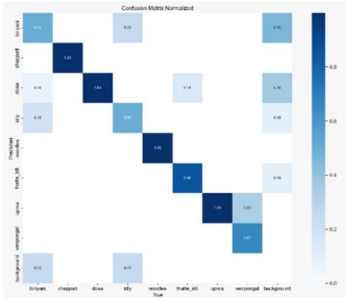

Figure 11. Confusion Matrix

The segmentation model (YOLO11) was evaluated based on multiple performance metrics:

- mAP (Mean Average Precision): 95.6%
- Precision: 85.6%
- Recall: 85.8%

The high mAP score of 95.6% demonstrates the model's strong capability to accurately segment food items, which is crucial for precise depth estimation and volume computation. The balance between precision and recall ensures that the model effectively identifies food regions while minimizing false positives and false negatives.

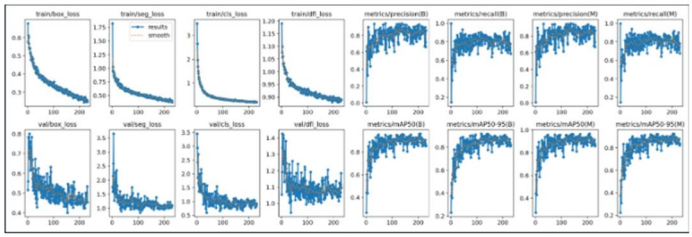

Figure 12. Performance graphs of YOLO11-seg model

Confusion Matrix: The model's performance was further evaluated by analyzing the confusion matrix as shown in Figure 11, which helped identify the locations of misclassifications. By examining the matrix, we pinpointed specific areas where the segmentation and classification models needed improvement, particularly for food items with intricate or similar attributes.
In conclusion, the final output of the model provides the amounts of proteins, carbohydrates, fats, fiber, calories, and sugar in the food, calculated based on the estimated weight.

VI. Conclusion and Future Work

This study presents a system that leverages advanced computer vision techniques for food classification, segmentation, and 3D volume estimation. Using YOLO v11 and InceptionV3, it achieved accuracies of 84.615% and 96.88%, respectively, while DepthAnythingV2 enabled precise volume and weight estimations, excelling with foods like biryani (93.89%) but facing challenges with complex textures like noodles (78.2%) and upma (59.6%). Future work will enhance depth estimation with multi-view imaging, optimize inference for mobile devices, expand dataset diversity, and improve weight accuracy through image-based food density estimation. This research has significant potential for dietary monitoring and personalized nutrition, particularly for the Indian demographic.

References

[1] H. Gao, Y. Liu, J. Li, and J. Gao, "Food nutrient extraction based on image recognition and entity extraction," Columbian College of Arts & Sciences, Phillips Hall, 2023, these authors contributed equally to this work.

[2] J. H. Kim, D. S. Lee, and S. K. Kwon, "Food classification and meal intake amount estimation through deep learning," Department of Computer Software Engineering, Dong-Eui University, 2023, correspondence: skkwon@deu.ac.kr; Tel.: +82-51-890-1727.

[3] Z. Shao, G. Vinod, J. He, and F. Zhu, "An end-to-end food portion estimation framework based on shape reconstruction from monocular image," in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), West Lafayette, Indiana, USA, 2023.

[4] H. Rahman, M. R. Pickering, D. A. Kerr, and E. J. Delp, "Food volume estimation in a mobile phone-based dietary assessment system," in Proceedings of SITIS 2012, 2012.

[5] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3D reconstruction for food volume estimation," in IEEE Conference Proceedings, 2023.

[6] A. Graikos, V. Charisis, D. Iakovakis, S. Hadjidimitriou, and L. Hadjileontiadis, "Single image-based food volume estimation using monocular depth-prediction networks," Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 2023.

[7] Z. S. E. F. Nandor Bandi, R. B. Tunyogi, and C. Sulyok, "Image-based volume estimation using stereo vision," IEEE 18th International Symposium on Intelligent Systems and Informatics, 2020. [Online]. Available: https://doi.org/10.1109/SWC50871.2021.00083

[8] N. Ratisoontorn and N. Chatwattanasiri, "Integrated image processing framework to determine nutrient quantities from guideline daily amount (GDA) label," IEEE SmartWorld Conference, 2021.

[9] T. Suzuki, K. Futatsuishi, K. Yokoyama, and N. Amaki, "Point cloud processing method for food volume estimation based on dish space," IEEE, 2020.

[10] F. S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "A review of image-based food recognition and volume estimation artificial intelligence systems," IEEE Reviews in Biomedical Engineering, vol. 17, 2024.

[11] J. Sultana, B. M. Ahmed, M. M. Masud, A. K. O. Huq, M. E. Ali, and M. Naznin, "A study on food value estimation from images: Taxonomies, datasets, and techniques," IEEE Access, vol. 11, 2023.

[12] S.-C. Huang, W.-C. Chiang, Y.-T. Yang, and J.-S. Wang, "An image-based AI nutrition analysis platform for food in compartment trays," IEEE International Congress on Advanced Applied Informatics, 2023.

[13] H. C. K. Y. Yu, Z. Yikun, and H. Xia, "A method of 3D point cloud volume calculation based on slice method," International Conference on Intelligent Control and Computer Application (ICCA), 2016.

[14] D. M. Ankit Basrur and A. R. Joshi, "Food recognition using transfer learning," IEEE International Conference on Industry Applications, 2022.

[15] N. S. T. Fotios S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "Weight estimation of Mediterranean food images using random forest regression algorithm," 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2023.

[16] K. F. Takuo Suzuki and K. Kobayashi, "Food volume estimation using 3D shape approximation for medication management support," 2018 3rd Asia-Pacific Conference on Intelligent Robot Systems, 2018.