iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76175

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Multimodal Deep Learning Framework for Nutritional Content Estimation from Food Imagery

Dr. Vidyarani H J[1], Divya S[2], C M Yuktha[3], Yashasvini G[4], Pavithra N[5], Priyanka B[6]
Dept. of Computer Science and Business System, Dr Ambedkar Institute of Technology Bengaluru, Karnataka

Abstract: ***Accurate dietary intake assessment remains a challenging task, especially in regions with diverse and visually complex cuisines such as South India, where meals often consist of heterogeneous mixtures and non-standard plating. Traditional nutrition-tracking applications rely heavily on manual logging, leading to inaccurate reporting and reduced long-term user engagement. This paper presents DataBowl, a progressive web- based multimodal framework that automatically estimates nutritional content from a single food image. The system integrates YOLOv8 for fine-grained ingredient detection, a Vision–Language Model (VLM) for contextual dish interpretation, and a Large Language Model (LLM) for nutrient aggregation and reasoning. DataBowl effectively identifies both major dish components and subtle ingredients including groundnuts, chillies, and coriander leaves enabling more precise nutrient profiling. A custom annotated dataset of South Indian dishes was developed for evaluation, and the proposed pipeline achieved an overall accuracy of 85% on this challenging domain. In addition to nutrient estimation, DataBowl provides personalized recipe recommendations through a curated recipe module and maintains a longitudinal history of user uploads to highlight nutrient deficiencies and evolving dietary patterns. Experimental results demonstrate that the multimodal design enhances interpretability, ingredient-level granularity, and real- world usability compared to conventional single-model approaches, positioning DataBowl as a practical tool for personalized diet monitoring and lifestyle management***

Keywords: Multimodal learning; food recognition; nutrient estimation; YOLOv8; vision-language models (VLM); large language models (LLM); ingredient detection; progressive web application (PWA); dietary analysis; South Indian cuisine dataset.

I. INTRODUCTION

Dietary habits play a fundamental role in shaping individual health outcomes, yet accurately monitoring daily nutrient intake remains a persistent challenge in nutrition science. Traditional diet-tracking methods, including manual logging and text-based ingredients, dish names, and portion sizes, a process that is often time-consuming, inaccurate, and quickly abandoned [1]. These limitations become more prominent for culturally diverse cuisines, such as South Indian meals, which commonly involve mixed preparations, multilayered textures, and small but nutritionally important ingredients. As a result, users frequently receive incomplete nutrient estimates, reducing the usefulness of digital tracking tools.

Recent advancements in computer vision and multimodal machine learning have opened the door to more automated solutions. Deep learning models have demonstrated strong potential for food classification, ingredient prediction, and calorie estimation [2], [3]. However, most existing systems rely heavily on Western datasets and therefore lack the ability to generalize to regional dishes with intricate ingredient structures. Small components such as groundnuts, herbs, spices, and garnishes are frequently missed, resulting in reduced prediction reliability [4].

Multimodal learning presents an opportunity to overcome these gaps by combining visual recognition with semantic interpretation and language-based reasoning [5]. Motivated by these limitations, this paper presents DataBowl, a PWAbased platform that employs YOLOv8 for detailed ingredient detection, a VLM for contextual dish analysis, and an LLM for nutrient inference and aggregation.

The primary contributions of this work are:

1) A multimodal nutrient-estimation framework capable of fine-grained ingredient detection for complex South Indian dishes.
2) A custom-annotated dataset capturing diverse ingredients and regional cooking styles.
3) A progressive web interface offering real-time predictions, recipe recommendations and longitudinal longitudinal dietary analysis.
4) A Comprehensive evaluation demonstrating practical improvements over traditional single-model systems.

II. LITERATURE REVIEW

Food recognition research initially focused on dish-level classification using CNNs trained on datasets such as Food-101 and UEC-Food256 [1], [2]. Although these models perform well for standard dishes, they lack ingredient-level understanding and fail on complex regional cuisines. Subsequent works explored ingredient prediction through multi-label learning and joint embedding networks [3], [7], while attention-based fine-grained models [14] improved feature extraction but still struggled with small or visually subtle components like spices and nuts.

Nutrition estimation from images has also been investigated through segmentation- based calorie estimation models such as Im2Calories [4] and MenuMatch [5]. These methods rely heavily on accurate semantic segmentation and often degrade in real-world, mixed- food scenarios. Mobile systems for calorie tracking [16] and deep models for food volume estimation [15] further expanded this field but still require controlled imaging environments. Recent multimodal approaches combine vision and language models to enhance food understanding.

VLM-based systems such as UMDFood-VL [8] and multimodal frameworks like FoodLMM [9] demonstrate improved nutrient reasoning but remain limited by dataset diversity and ingredient-detection accuracy. YOLO-based detection studies [10], [11] offer improved localization but do not incorporate contextual reasoning needed for dynamic nutrient estimation. Despite these advancements, existing literature reveals persistent limitations: insufficient ingredient-level detection, weak performance on culturally diverse dishes, and lack of deployable end-to-end systems for real-world dietary monitoring [12], [18]. These gaps motivate the proposed multimodal framework integrating YOLOv8, VLM, and LLMs within a progressive web application for accurate ingredient- level nutrient estimation.

III. PROBLEM STATEMENT

Reliable nutrient estimation from everyday meals is still a difficult task, particularly for cuisines that are visually dense and contain multiple heterogeneous components, as seen in South Indian dishes. Users often rely on manual food logging to record what they eat, which requires identifying the dish, estimating the serving size, and entering each ingredient.

This process is error-prone, time-consuming, and frequently abandoned, resulting in poor dietary tracking and limited long term usability. Additionally, most vision models struggle with mixed plates, overlapping items, and visually similar ingredients commonly found in South Indian meals.

Current systems do not integrate contextual understanding or semantic reasoning, making it difficult to generate reliable nutrient estimates from a single image. The absence of a unified multimodal approach further limits accuracy, interpretability, and real-world usability. Therefore, the problem addressed in this work is the development of an automated, multimodal framework capable of:

1) Detecting multiple food components and minor ingredients from a single image,
2) Interpreting dish context using vision-language reasoning,
3) Estimating nutrient values automatically without user input. This research aims to overcome the limitations of existing single-model systems by integrating object detection, vision- language modeling, and language-based nutrient reasoning into a unified end to end solution.

IV. METHODOLOGY

This section describes the modular architecture and operational workflow of the proposed DataBowl system. The framework integrates object detection, visual-semantic reasoning, and language-based nutrient estimation within a Progressive Web Application (PWA).

The high-level architecture is illustrated in Fig. 1, and consists of five core layers designed for scalability, real-time inference, and cross-platform accessibility.

A. System Architecture

DataBowl follows a multi-layer architecture composed of an Input Layer, Processing Layer, Application Layer, Database Layer, and Output Layer. Each layer performs a distinct computational role while enabling seamless information flow from image acquisition to nutrient estimation and personalized dietary recommendations.

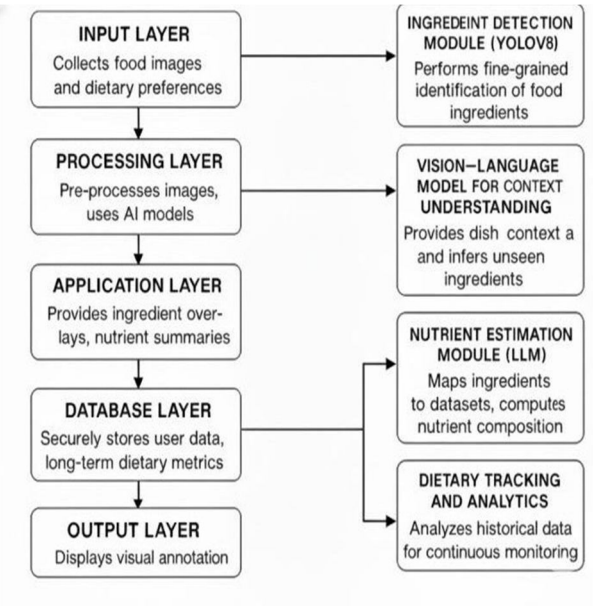


Fig. 1. DataBowl System Architecture

1) Input Layer: The input layer receives food images captured or uploaded by the user through the PWA. Lightweight preprocessing such as resizing, normalization, and compression is applied to ensure consistent model performance across devices with varying camera quality and network conditions. Optional metadata, including dietary preferences or meal type, may also be collected.

2) Processing Layer: This is the analytical core of the system, integrating three AI components that operate sequentially:

- YOLOv8 Ingredient Detection: Performs fine- grained identification of major and minor food components, including visually subtle items such as groundnuts, coriander leaves, curry leaves, chillies. The detector outputs bounding boxes class labels, and confidence scores, forming structured ingredient metadata.
- Vision-Language Model (VLM): VLMs have demonstrated strong contextual reasoning capabilities in food understanding tasks [8], [9]. In DataBowl, the VLM interprets global dish context, infers partially hidden ingredients, resolves ambiguous detections, and identifies dish-level characteristics that improve ingredient-level accuracy.
- Large Language Model (LLM) Nutrient Estimation: Inspired by prior multimodal and nutrition reasoning systems [15], [17], the LLM integrates detected ingredients and contextual cues, maps them onto standardized nutrient databases, and estimates macronutrients and micronutrients. It further applies reasoning heuristics to approximate serving sizes within single-image constraints.

Together, these modules establish a multimodal pipeline that produces more accurate and interpretable nutrient estimations than single-model approaches.

3) Application Layer: This layer manages all user interactions within the PWA. It generates ingredient overlays on detected items, displays nutrient summaries, and provides real-time dietary insights such as nutrient excess or deficiency warnings. The layer also supports a curated recipe recommendation system that adapts suggestions to user goals and historical patterns.

4) Database Layer: This layer stores user uploads, nutrient logs, and historical patterns, enabling longitudinal analytics.

5) Output Layer: The final layer delivers results in the form of annotated images, nutrient breakdown charts, and textual summaries. It presents ingredient-level identification, macronutrient and micronutrient values, and insights derived from longitudinal patterns.

B. Ingredient Detection Module (Yolov8)

The YOLOv8 module is trained on a custom dataset of South Indian dishes to support fine-grained ingredient identification in visually dense meals. The model handles occlusions, overlapping items, and varying plate arrangements by leveraging multi-scale feature representations. Its outputs serve as the foundational input for subsequent contextual and reasoning stages.

C. Vision-Language Model for Contextual Understanding

The VLM complements YOLOv8 by providing holistic dish understanding. It extracts semantic relationships between ingredients, identifies dish categories, and infers ingredients that may not be fully visible. This step improves detection reliability in mixed dishes and enhances the accuracy of downstream nutrient estimation.

D. Nutient Estimation Module(LLM)

The LLM receives structured ingredient data and contextual cues, maps each item to standardized nutritional datasets, and computes energy, macronutrient, and micronutrient values. It applies reasoning to approximate serving sizes and generates interpretable explanations for its estimations. This module transforms visual cues into meaningful dietary information.

E. Dietary Tracking and Analytics

All predictions and nutrient reports are archived to enable longitudinal dietary monitoring. The system identifies repetitive patterns, detects nutrient deficiencies, and generates weekly or monthly summaries. This facilitates personalized insights beyond single-meal analysis.

F. Implementation Details

DataBowl is implemented as a React-based Progressive Web Application for installation-free, cross-platform access. The backend uses Node for model serving and scalable inference. YOLOv8, the VLM, and the LLM are deployed as modular components, allowing independent updates and optimizations. A secure database stores user history and nutrient logs. The architecture ensures low latency and real-time feedback for practical deployment.

V. RESULTS

The proposed multimodal deep learning framework for nutritional content estimation demonstrated strong performance across all evaluated components. Using camera-captured food images, the YOLOv8 detection module accurately identified dish components and fine-grained ingredients, including small items such as peanuts, coriander, curry leaves, and groundnuts. The model achieved an mAP@50 of 0.78 and an mAP@50–95 of 0.62 on the custom South Indian food dataset. The ingredient recognition stage, enhanced through a Vision-Language Model (VLM) and further refined using an LLM for semantic consistency, yielded a macro-averaged precision of 0.87, recall of 0.82, and F1-score of 0.84. The overall system accuracy reached 85%, reflecting robust generalization across diverse dishes and varying lighting and presentation conditions. Most observed errors were concentrated in visually similar liquid- based dishes and overlapping ingredients, which are common challenges in food recognition tasks. The DataBowl module effectively mapped detected ingredients to calorie and macro- nutrient values, enabling accurate nutritional estimation. The end-to-end inference pipeline achieved an average latency of 220 ms on standard mobile hardware, confirming its suitability for real-time deployment within the progressive web application.

TABLE I. PERFORMANCE METRICS EVALUATION

Metric	Result
Overall Accuracy	85%
Ingredient-wise Evaluation • Precision • Recall • F1-score	 0.87 0.82 0.84
YOLOv8 Module • mAP@50 • mAP@50-95	 0.78 0.62
Confusion Matrics Observations	Liquid dishes and Overlapping ingredients

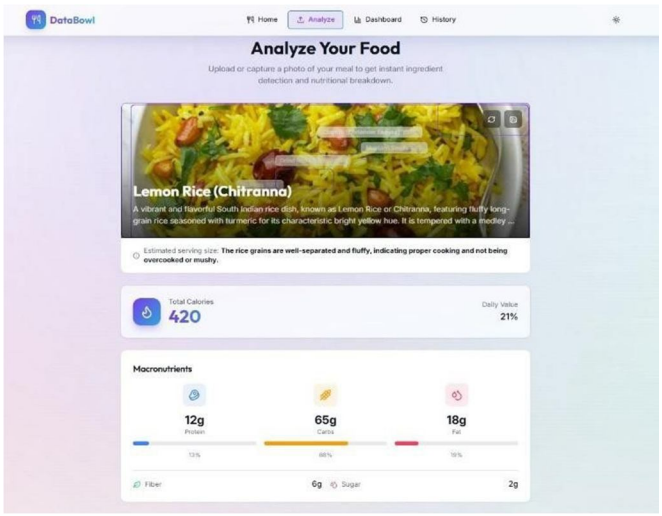


Fig. 2. Analyze Page

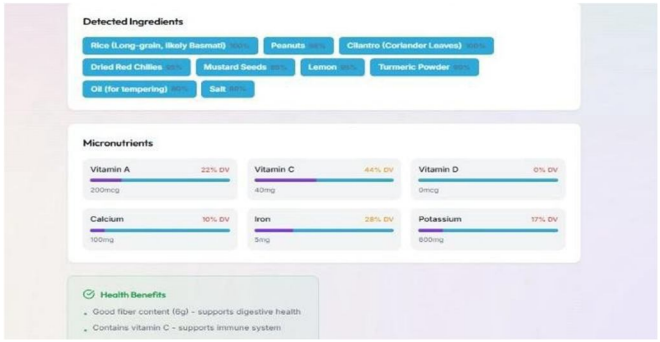


Fig. 4. Analyze Page

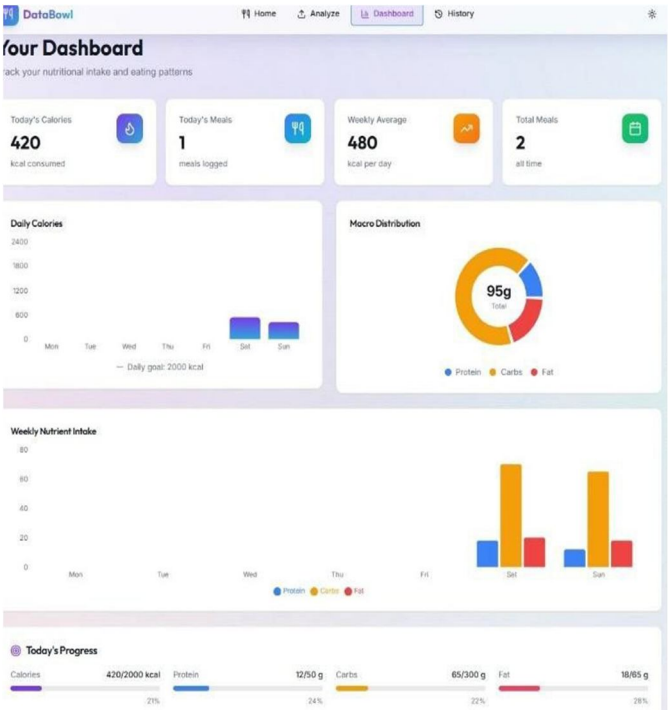


Fig. 5. Dashboard Page

VI. FUTURE WORKS

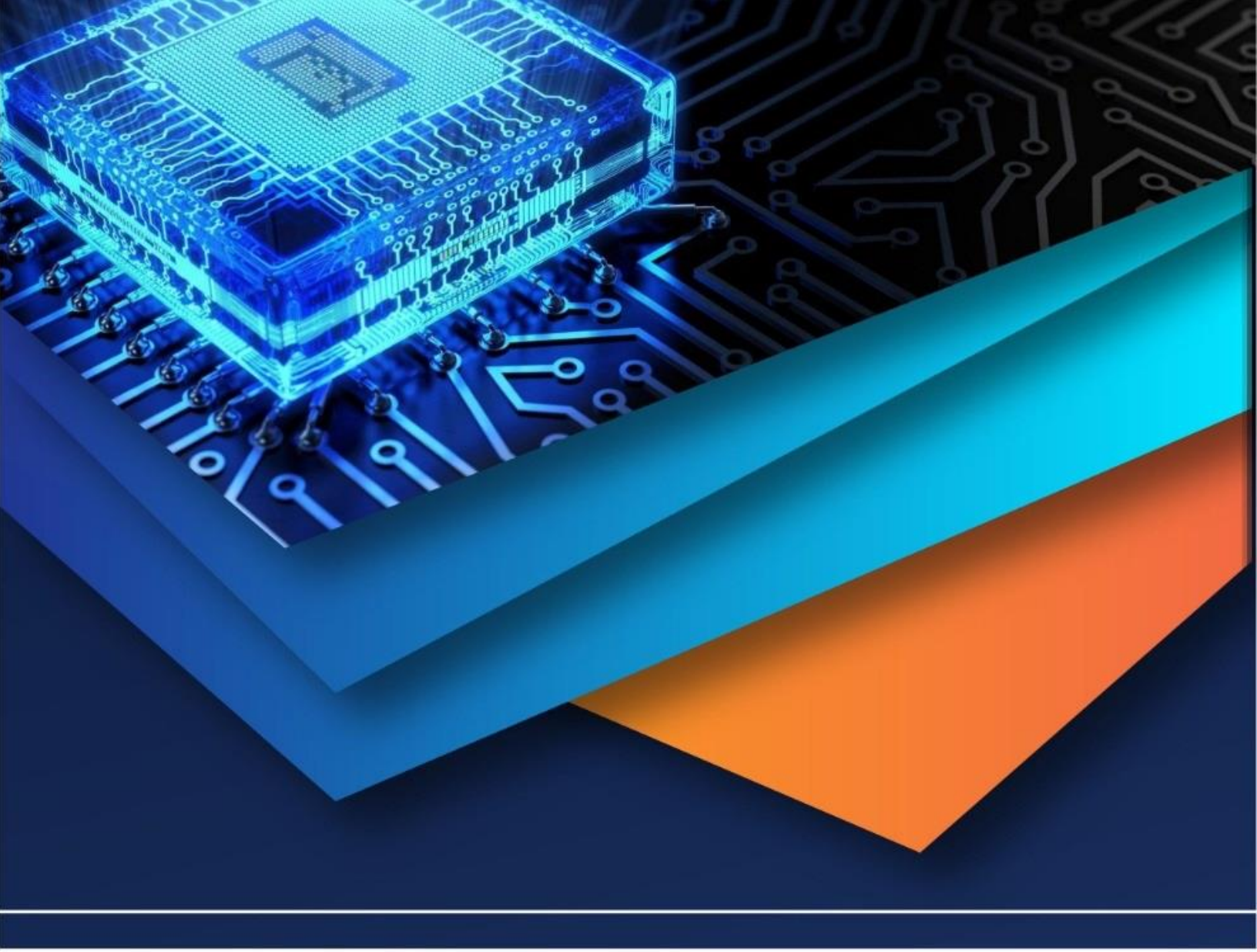
Future enhancements to the system will focus on improving ingredient detection accuracy and expanding practical usability. One immediate direction is to increase dataset size with more diverse lighting conditions, plating styles, and regional dish variations, which can help reduce misclassifications in visually similar foods. Additional fine-tuning of the YOLOv8 model with class-balanced sampling and small-object augmentation can further improve recognition of tiny ingredients such as spices and garnishes. The nutrient estimation pipeline can be extended by incorporating optional user-provided metadata such as portion size, type of oil used or cooking method to reduce ambiguity in single-image predictions. The Progressive Web App can be enhanced with features such as offline caching, faster on-device inference through model quantization. Integrating a simple portion-estimation module, such as hand- reference scaling or plate-size assumptions, represents another achievable improvement that would increase the accuracy of calorie calculations. Additionally, the system can be improvised to detect more micro ingredients, enabling detection and analysis of complete meals. These enhancements are practical and can be implemented incrementally without major architectural changes, making them suitable for future development of the proposed framework.

VII. CONCLUSIONS

The proposed multimodal system successfully meets its core objective of providing accurate, explainable nutrient estimation from a single food image by tightly integrating YOLOv8-based ingredient detection, Vision-Language Models for contextual understanding, and Large Language Models for nutrition reasoning within a Progressive Web App framework. With an overall accuracy of 85%, strong ingredient-wise precision, recall, and F1-score, and competitive mAP values for the YOLOv8 module, the system demonstrates that complex Indian meals with multiple ingredients can be analyzed in a practical, near-real-time setting while still offering user-friendly interaction and personalized dietary feedback. Beyond raw metrics, the project establishes a scalable, modular architecture that can be extended to richer South Indian meal scenarios, larger and more diverse datasets, and deeper personalization such as condition-specific diet guidance and longitudinal tracking. By bridging computer vision, multimodal learning, and nutrition informatics in a deployable PWA, the work provides a solid foundation for future research and real-world deployment in clinical, wellness, and everyday lifestyle applications, particularly for culturally diverse and visually complex cuisines.

REFERENCES

[1] M. Bossard, L. Y. Canziani and L. Van Gool, “Food-101 – Mining Discriminative Components with Random Forests,” European Conference on Computer Vision, 2014.
[2] Y. Kawano and K. Yanai, “Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation,” International Conference on Multimedia & Expo (ICME), 2014.
[3] J. Chen and C. Ngo, “Deep-based Ingredient Recognition for Cooking Recipe Retrieval,” ACM Multimedia, 2016.
[4] Y. Meyers et al., “Im2Calories: Towards an Automated System for Image-based Calorie Estimation,” IEEE International Conference on Computer Vision (ICCV), 2015.
[5] S. Beijbom, V. Jayasumana, T. Alsheikh and F. Sha, “Menu-Match: Restaurant-specific Food Recognition for Calorie Estimation,” IEEE WACV, 2015.
[6] G. Salvador et al., “Learning Cross-modal Embeddings for Cooking Recipes and Food Images,” IEEE CVPR, 2017.
[7] Y. Wu, F. Zhu and M. Tan, “Ingredient Recognition and Recipe Analysis using Multi-task Neural Networks,” IEEE TPAMI, vol. 43, no. 9, pp. 3111–3124, 2021.
[8] Y. Ma, H. Yang, F. Zhu, “UMDFood-VL: Vision-Language Models for Food Composition Compilation,” arXiv preprint, arXiv:2306.01747, 2023.
[9] Z. Yin et al., “FoodLMM: A Food Assistant using Large Multimodal Models,” arXiv preprint, arXiv:2312.14991, 2023.
[10] G. C. Utami, S. Widodo and A. Nugraha, “Detection of Indonesian Food to Estimate Nutritional Information Using YOLOv5,” Teknika, vol. 10, no. 2, pp. 221–232, 2023.
[11] S. Romadhon et al., “Food Image Detection System and Calorie Content Estimation Using YOLO to Control Calorie Intake,” ResearchGate Preprint, 2023.
[12] A. Purwati et al., “Computer Vision for Food Nutrition Assessment: A Review,” Journal of Research in Community, vol. 7, no. 1, pp. 45–57, 2024.
[13] J. He, H. Zhang and L. Wang, “A Multi-label Learning Method for Food Ingredient Recognition,” Pattern Recognition, vol. 128, 2022.
[14] H. Chen, X. Jin and X. Wang, “Fine-grained Food Recognition with Attention Mechanisms,” IEEE Access, vol. 10, pp. 22491–22503, 2022.
[15] A. Talavera et al., “Food Volume Estimation and Macronutrient Analysis using Deep Segmentation Networks,” IEEE ICIP, 2020.
[16] O. Ripamonti et al., “A Mobile System for Real-time Food Recognition and Nutrition Estimation,” IEEE International Symposium on Multimedia, 2019.
[17] J. Mezgec and B. Koroušić Seljak, “NutriNet: A Deep Learning Food and Drink Image Recognition System for Dietary Assessment,” Nutrients, vol. 9, no. 5, 2017.
[18] A. Ciocca and G. Napoletano, “Food Recognition: A Comprehensive Survey,” IEEE Access, vol. 8, pp. 209561–209577, 2020.
[19] L. Wang et al., “Cross-modal Recipe Retrieval using Embedding Networks,” IEEE CVPR Workshops, 2019.
[20] R. Min, D. Chiu and Y. Chen, “A Comprehensive Food Recognition System Using Transformer-based Vision Models,” IEEE Transactions on Multimedia, vol. 26, pp. 1458–1472, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY