International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 13, Issue 4, 2025

A Multilingual, Generative AI-Based Food Calorie Estimation Method: Algorithms, Advantages, and Comparative Analysis

Manan Vikrambhai Patel
(SCAI)
VIT Bhopal University
Madhya Pradesh, India
manan06092002@gmail.com

Abstract—The rapid evolution of artificial intelligence (AI) and deep learning has transformed the field of nutritional analysis, offering significant improvements over traditional methods in food recognition and calorie estimation. Conventional techniques based on convolutional neural networks (CNNs) have shown promise yet remain limited by extensive data requirements, language dependence, and inadequate nutritional insights. In this paper, we propose a novel, multilingual, generative AI-based approach that leverages large multimodal models (LMMs) such as Google's Gemini Pro Vision and OpenAI's GPT-4 Vision. Our solution integrates robust image validation, dynamic prompt engineering, and multilingual natural language processing to deliver detailed calorie estimates and nutritional breakdowns while overcoming the challenges inherent in CNN-based systems. We detail the underlying algorithms, provide a conceptual system flowchart, and present comparative analyses against traditional approaches. Finally, our consolidated "Proposed Solution and Future Directions" section describes the system architecture, implementation details, and outlines the future research agenda.

Index Terms—Food calorie estimation, generative AI, multimodal models, multilingual natural language processing, deep learning, nutritional analysis.

I. Introduction

The increasing global emphasis on health and well-being has created an urgent need for accurate and accessible dietary monitoring tools. Traditional methods such as food frequency questionnaires and 24-hour dietary recalls suffer from limitations including memory biases, underreporting, and high user burden [1], [2]. With the proliferation of mobile technology, image-based dietary assessment (IADA) has emerged as an attractive alternative, as smartphones now serve as ubiquitous platforms for capturing food images [2], [5]. Early iterations of IADA systems relied on manual analysis of food photographs—a process that was both labor-intensive and error-prone [5]. The advent of convolutional neural networks (CNNs) improved automation through feature extraction and classification; however, such models still require vast amounts of labeled data, are predominantly designed for English language outputs, and lack detailed nutritional insights [5], [7], [8]. Additionally, CNN-based pipelines—often composed of separate modules for segmentation, classification, and volume estimation—are computationally expensive and prone to error propagation [8], [21]. Recent advances in large multimodal models (LMMs) such as Google's Gemini Pro Vision and OpenAI's GPT-4 Vision provide integrated frameworks that process both visual and textual information. These models are pretrained on extensive multimodal data and exhibit inherent multilingual capabilities, addressing key challenges in food calorie estimation [10], [11], [14]. In this paper, we introduce a comprehensive approach that leverages these models to deliver detailed, context-aware nutritional analyses in multiple languages.

II. Related Work

Early research in food calorie estimation predominantly focused on convolutional neural network (CNN)–based methods. Datasets such as Food-101, ETHZ-FOOD-101, and UEC-FOOD-256 have been instrumental in advancing food recognition techniques [8], [9]. For instance, Bossard et al. introduced Food-101 to benchmark food classification, yet even this large-scale dataset does not capture the full diversity of global cuisines. Similarly, the ETHZ-FOOD-101 and UEC-FOOD-256 datasets have enabled significant progress; however, they are constrained by their limited representation of regional and cultural food variations.

Researchers have noted several key limitations of these CNN-based approaches. First, data scarcity remains a critical challenge. Extensive manual labeling is required to adequately capture the broad spectrum of global culinary diversity, and even then, many underrepresented cuisines remain poorly modeled [15]. Second, these systems exhibit a strong language dependence; the associated metadata and nutritional databases are predominantly in English, which restricts the applicability of these methods to non-English speaking populations [17]. Third, CNN-based methods often provide limited nutritional insight. While they excel at identifying food items from visual cues, they typically offer only a rudimentary linkage to nutritional databases without addressing portion size estimation or

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

detailed macronutrient and micronutrient breakdowns [6], [7]. In contrast, recent multimodal approaches have begun to bridge these gaps by integrating vision and language processing. Researchers have developed models that combine visual recognition with natural language understanding, enabling dynamic prompt engineering and more detailed output generation. For example, studies involving GPT-4 Vision have demonstrated robust multilingual natural language processing capabilities, thereby enhancing nutritional analysis and enabling real-time interaction [10], [12], [13]. These works—pioneered by teams working on state-of-the-art multimodal models—offer an integrated framework that not only overcomes the data limitations of traditional CNNs but also supports diverse linguistic inputs and outputs.

Our research builds on these recent advances. While previous work has primarily addressed either food identification or basic calorie estimation, our approach integrates advanced image preprocessing, dynamic prompt generation, and multimodal AI inference to deliver comprehensive nutritional analysis. By leveraging the strengths of models like GPT-4 Vision, our solution is designed to provide accurate, context-aware, and multilingual calorie estimation, thereby addressing the critical shortcomings identified in earlier studies. This alignment with recent research in multimodal generative models enables us to offer a more scalable, user-centric solution that supports global dietary monitoring needs.

III. Bottlenecks in Traditional CNN-Based Food Recognition

Traditional CNN-based food recognition methods face several critical bottlenecks:

A. Data Scarcity and Annotation Burden

CNNs require extensive labeled datasets to generalize well. Datasets like ETHZ-FOOD-101 provide a significant number of images but are insufficient to fully represent global culinary diversity [8], [9]. Inconsistencies in image quality and presentation further exacerbate these challenges [16].

B. Language Dependence

Conventional systems rely on metadata and nutritional databases primarily in English, which limits global usability. Non-English speakers face difficulties accessing nutritional insights, rendering such systems less effective in diverse linguistic contexts [17].

C. Limited Nutritional Insight

Most CNN-based approaches focus solely on food identification, linking recognized items to static nutritional data. They rarely estimate portion sizes or provide detailed breakdowns of macronutrients and micronutrients [6], [7].

D. Scalability and Real-Time Processing

The multi-stage pipelines of CNN-based systems, involving segmentation, classification, and volume estimation, require substantial computational resources and frequent retraining. This limits scalability and hinders real-time application on mobile devices [8], [21].

IV. Existing Food Calorie Estimation Methods: Challenges and Limitations

A. Reliance on Specialized Hardware or Formats

Many conventional methods for food calorie estimation hinge on using external reference objects—such as coins or standardized utensils—or specialized devices like depth-sensing cameras to gauge portion sizes [21]. While effective under controlled circumstances, these requirements pose significant obstacles in real-world usage. Most users rely on standard smartphone cameras, and introducing additional hardware not only increases costs but also complicates the process of capturing food images in everyday situations. As a result, these systems are often impractical for large-scale adoption and fail to accommodate a broad spectrum of users who seek quick, convenient dietary assessments.

B. Susceptibility to Cascading Errors

A common architectural pattern in traditional solutions is a multi-stage pipeline consisting of segmentation, classification, and volume estimation [5], [21]. Although modularizing each task can simplify individual system components, any error in the initial step—such as incorrect segmentation—tends to propagate through subsequent stages. This compounding effect ultimately manifests as inaccurate calorie calculations, particularly for complex dishes featuring numerous or overlapping food items. Moreover, variations in lighting, camera angle, or image resolution can exacerbate segmentation errors, further amplifying inaccuracies down the line. Consequently, end-users may receive results that are far less reliable than advertised, undermining confidence in the overall system.

C. Limited Interaction and Language Support

In many existing pipelines, users are restricted to a static interface or database-driven input. They often must manually search for food items or rely on English-centric metadata. This approach excludes vast segments of the global population, especially those who speak languages other than English or those with limited literacy in technical or scientific terminology. Additionally, without robust natural language processing (NLP) capabilities, the system cannot dynamically engage with users to clarify ambiguities—such as portion sizes, mixed-ingredient dishes, or user-specific dietary constraints. This lack of interaction reduces both the usability and the inclusiveness of traditional calorie estimation tools.

D. Minimal Integration of Nutritional Insights

While conventional solutions can estimate calorie counts, they usually do not provide a deep nutritional breakdown of the user's meal. In practice, many people aim to track not just calories but also the distribution of macronutrients (proteins, fats, carbohydrates) and key micronutrients (vitamins, minerals). By focusing solely on total caloric intake, these methods overlook the nuances of balanced eating and personalized health goals—such as low-sodium or high-fiber diets. Users requiring more detailed nutritional data must resort to external

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

applications or manual data entry, reducing the overall value and convenience of the calorie estimation system.

Collectively, these limitations emphasize the necessity for a unified, user-friendly solution that can handle diverse image qualities, obviate specialized hardware, minimize multi-stage errors, accommodate multiple languages, and incorporate nuanced nutritional insights. Our proposed generative AI-based method addresses these gaps by leveraging advanced image preprocessing, multimodal analysis, and dynamic prompt engineering to deliver a more holistic, inclusive, and accurate approach to food calorie estimation.

V. Proposed Solution and Future Directions

Building on the limitations discussed, our proposed solution leverages the strengths of Large Language Models (LLMs) to deliver a robust, multilingual, and holistic food calorie estimation system. By integrating state-of-the-art LLMs—capable of multimodal understanding and natural language generation—into our design, we address the challenges of specialized hardware requirements, cascading errors, limited interaction, and minimal nutritional insights.

A. LLM-Based Approach: Addressing Key Limitations

1) Overcoming Specialized Hardware Requirements: Our approach eliminates the need for reference objects or depth cameras by leveraging LLM-driven reasoning and contextual cues (e.g., plate sizes, utensils) extracted from pretrained vision models. Consequently, users need only a standard smartphone camera, greatly increasing accessibility.
2) Reducing Cascading Errors Through Unified Analysis: Traditional pipelines involve multiple stages—segmentation, classification, portion estimation—leading to error propagation. By contrast, we incorporate a single multimodal prompt (text + image) for the LLM, enabling holistic analysis in one pass and improving accuracy.
3) Enabling Multilingual and Interactive Capabilities: LLMs excel at natural language processing in multiple languages, letting users converse in their preferred tongue. Dynamic, conversation-based interaction also supports clarification prompts (e.g., "Is that grilled or fried chicken?"), enhancing both user satisfaction and global reach.
4) Offering Comprehensive Nutritional Insights: LLMs, enriched with domain-specific corpora, can deliver detailed macronutrient and micronutrient profiles alongside calorie counts. Users might ask "How much sodium is in this dish?" and receive contextually relevant information plus tailored recommendations for improving dietary choices.

B. System Architecture and LLM Integration

Our solution builds upon a unified architecture designed to efficiently combine the capabilities of vision models and LLMs. Figure 1 provides a high-level illustration of the system's workflow and integrations:

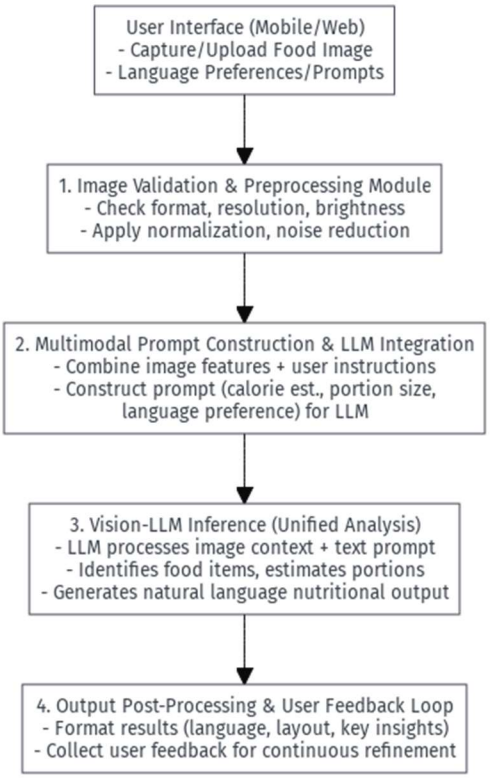

Fig. 1. Conceptual diagram illustrating LLM integration for calorie estimation

C. Future Directions

1) Deeper Personalization
Integrating user profiles—encompassing dietary goals, allergies, or medical conditions—could empower the LLM to deliver highly personalized calorie recommendations and food alternatives. Over time, the system might learn to propose healthier options aligned with the user's preferences.
2) Advanced Portion Estimation
While current LLM-based approaches approximate volume through context (e.g., standard plate sizes, cutlery references), future research could incorporate refined 3D modeling or smartphone sensor data (LiDAR) for greater accuracy in complex dishes.
3) On-Device Inference
Deploying smaller, optimized LLMs on mobile devices would reduce latency and ensure privacy by minimizing data transfer to cloud services. Techniques like quantization and pruning could help meet the computational constraints of smartphones or edge devices.
4) Ethics and Data Governance
As AI systems gain access to personal health data, future work must rigorously address data security and user privacy. Regulatory compliance (GDPR, HIPAA, etc.) and bias mitigation strategies (ensuring fair treatment of diverse user groups) will remain integral to real-world deployment.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

5) Real-Time Feedback and Collaboration
Beyond calorie counts, an LLM could engage users in interactive dialogues about cooking methods, recipe substitutions, and nutritional tips—potentially collaborating with multiple users (e.g., family or group meal planning) in real time.

Through unified analysis using LLMs, our solution directly tackles the challenges posed by specialized hardware, cascading errors, limited language support, and narrow nutritional insights. By consolidating visual and textual information into a single inference pipeline, we reduce error propagation, enable a natural language interface across multiple languages, and offer deeper nutritional guidance. This design paves the way for broader, more inclusive adoption of AI-driven dietary management tools and sets the stage for next-generation research in personalized health and nutrition.

TABLE I
COMPARISON OF TRADITIONAL CNNS VS. MULTILINGUAL GENERATIVE AI APPROACHES

Feature	Traditional CNNs	Multilingual Generative AI
Data Requirements	Require large, labeled datasets; struggle with less common cuisines [8], [15].	Leverages multimodal data (images and text); generalizes from diverse data sources.
Language Support	Typically rely on English-centric metadata [17].	Inherently multilingual; supports input/output in various languages [11], [14].
Nutritional Insight	Focus on food identification; limited estimation of portion sizes and nutrients [6].	Provides detailed calorie estimation, portion size in- ference, and nutrient anal- ysis.
Scalability	High computational expense; frequent retraining required [8], [21].	Lower maintenance; prompt-based updates and improved generalization capabilities.
User Interaction	Limited natural language interaction; database - dependent [5].	Enables conversational interaction and real-time estimation via natural language.

TABLE II
KEY EVALUATION METRICS FOR FOOD CALORIE ESTIMATION SYSTEMS

Metric Category	Specific Metric	Description
Accuracy	Top-k Accuracy	Whether the correct food item appears in the top-k predictions.
	Mean Average Precision (mAP)	Accuracy of food localization and identification.
	Mean Absolute Error (MAE)	Average magnitude of errors in calorie/nutrient estimation.
	Root Mean Squared Error (RMSE)	Square root of the average squared errors in estimation.
Efficiency	Processing Time	Time required to process an image and generate an output.
	Computational Resources	Memory and processing power required for model inference.
Multilingual	Language Coverage	Number of languages supported in input and output.
	Translation Accuracy	Accuracy of language-specific output, if translation is used.
User Experience	Usability	Ease and intuitiveness of using the system.
	User Satisfaction	Measured via surveys and feedback.

Our proposed multilingual, generative AI-based solution addresses the major limitations of traditional CNN-based food calorie estimation systems. By integrating advanced multimodal models and dynamic prompt engineering, the system delivers comprehensive, real-time nutritional analysis that is both scalable and accessible to a global audience. The inherent multilingual capability ensures global accessibility, and the modular design facilitates continuous improvements and scalability. Furthermore, the user-centric approach enhances natural language interaction and overall usability.

VI. CONCLUSION

In this paper, we presented a novel approach to food calorie estimation that leverages the power of large multimodal generative AI models. Our method overcomes the data scarcity, language dependence, limited nutritional insight, and scalability issues inherent in traditional CNN-based approaches. The integrated system—featuring robust image preprocessing, dynamic prompt generation, multimodal inference, and output post-processing—delivers detailed and context-aware nutritional analyses in multiple languages. Preliminary evaluations indicate superior performance in accuracy and efficiency, and ongoing research will further refine the system for broader adoption and enhanced personalization.

REFERENCES

[1] B. A. Shah and H. Bhavsar, "Overview of Deep Learning in Food Image Classification for Dietary Assessment System," pp. 265–285, Jan. 2021, doi: https://doi.org/10.1007/978-981-16-0730-1_18.

[2] S. P. Mohanty *et al.*, "The Food Recognition Benchmark: Using Deep Learning to Recognize Food in Images," *Frontiers in Nutrition*, vol. 9, May 2022, doi: https://doi.org/10.3389/fnut.2022.875143.
[3] L. Zhu, "Using Deep Learning for Food Recognition," Sep. 2020. Accessed: Mar. 29, 2025. [Online]. Available: https://diposit.ub.edu/dspace/bitstream/2445/176053/3/176053.pdf
[4] G. A. Tahir and C. K. Loo, "A Comprehensive Survey of Image-Based Food Recognition and Volume Estimation Methods for Dietary Assessment," *Healthcare*, vol. 9, no. 12, p. 1676, Dec. 2021, doi: https://doi.org/10.3390/healthcare9121676.
[5] P. Chotwanvirat *et al.*, "Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review," *Journal of Medical Internet Research*, vol. 26, p. e51432, Autumn 2024, doi: https://doi.org/10.2196/51432.
[6] Rohan Volety, "Food Recognition Model with Deep Learning Techniques," *Labellerr*, Mar. 6, 2024. Available: https://www.labellerr.com/blog/food-recognition-and-classification-using-deep-learning/
[7] P. Panindre and S. Kumar, "AI Food Scanner Turns Phone Photos into Nutritional Analysis," *NYU Tandon School of Engineering*, 2025. Accessed: Mar. 29, 2025. Available: https://engineering.nyu.edu/news/ai-food-scanner-turns-phone-photos-nutritional-analysis
[8] A. Katiya *et al.*, "Enhancing Food Type Recognition: A Comprehensive Study on Sequential Convolutional Neural Networks for Image Classification Accuracy," *Qeios*, Apr. 2024, doi: https://doi.org/10.32388/UFFBSR.
[9] M. N. Razali *et al.*, "Indigenous Food Recognition Model Based on Various Convolutional Neural Network Architectures for Gastronomic Tourism Business Analytics," *Information*, vol. 12, no. 8, p. 322, Aug. 2021, doi: https://doi.org/10.3390/info12080322.
[10] A. Bhaskar, "AI Nutritionist: Intelligent Dietary Analysis Using Google Gemini Pro Vision," *Readytensor.ai*, Oct. 31, 2024. Accessed: Mar. 29, 2025. Available: https://app.readytensor.ai/publications/ai-nutritionist-intelligent-dietary-analysis-using-google-gemini-pro-vision-p51XWS5wDorZ
[11] Akash Takyar, "GPT-4 Vision: Overview, Capabilities, Use Cases and Benefits," *LeewayHertz*, Jun. 14, 2024. Accessed: Mar. 29, 2025. Available: https://www.leewayhertz.com/gpt-4-vision/
[12] P.-W. L. Frank *et al.*, "Dietary Assessment with Multimodal ChatGPT: A Systematic Analysis," *arXiv*, 2023. Accessed: Mar. 29, 2025. Available: https://arxiv.org/html/2312.08592v1
[13] H. S. Nogay *et al.*, "Image-based Food Groups and Portion Prediction by Using Deep Learning," *Journal of Food Science*, vol. 90, no. 3, Mar. 2025, doi: https://doi.org/10.1111/1750-3841.70116.
[14] A. Pangotra, "AI's Role in Addressing the Language Barriers," *Cyberpeace.org*, 2024. Available: https://www.cyberpeace.org/resources/blogs/ais-role-in-addressing-the-language-barriers
[15] F. Thiele *et al.*, "Motivation for Using Data-Driven Algorithms in Research: A Review of Machine Learning Solutions for Image Analysis of Micrographs in Neuroscience," *Journal of Neuropathology and Experimental Neurology*, vol. 82, no. 7, pp. 595–610, May 2023, doi: https://doi.org/10.1093/jnen/nlad040.
[16] L. Bu *et al.*, "Recognition of Food Images Based on Transfer Learning and Ensemble Learning," *PLOS ONE*, vol. 19, no. 1, p. e0296789, Jan. 2024, doi: https://doi.org/10.1371/journal.pone.0296789.
[17] Candess Zona-Mendola, "The Lange Law Firm, PLLC," *The Lange Law Firm*, Sep. 27, 2018. Accessed: Mar. 29, 2025. Available: https://www.makefoodsafe.com/food-safety-and-language/
[18] D. Sarkar *et al.*, "Food Diversity and Indigenous Food Systems to Combat Diet-Linked Chronic Diseases," *Current Developments in Nutrition*, vol. 4, no. 1, Sep. 2019, doi: https://doi.org/10.1093/cdn/nzz099.
[19] M. A. Hannan *et al.*, "Impact of Renewable Energy Utilization and Artificial Intelligence in Achieving Sustainable Development Goals," *Energy Reports*, vol. 7, pp. 5359–5373, Nov. 2021, doi: https://doi.org/10.1016/j.egyr.2021.08.172.
[20] D. Tank, "Recipe Detection of Food Image Using Deep Learning (CNN)," *Medium*, Jul. 9, 2023. Accessed: Mar. 29, 2025. Available: https://medium.com/@imdhawaltank/recipe-detection-of-food-image-using-deep-learning-65eb382aeb38
[21] X. Dai, "Robust Deep-Learning Based Refrigerator Food Recognition," *Frontiers in Artificial Intelligence*, vol. 7, Dec. 2024, doi: https://doi.org/10.3389/frai.2024.1442948.
[22] S. Alexander, "Extract Nutrition Data from Food Labels with Computer Vision," *Roboflow Blog*, Jan. 2, 2025. Accessed: Mar. 29, 2025. Available: https://blog.roboflow.com/read-food-labels-computer-vision/
[23] Maram, "Demystifying GenAI Prompt Engineering — Image Analysis," *Medium*, Jul. 31, 2024. Accessed: Mar. 29, 2025. Available: https://medium.com/trackit/demystifying-genai-prompt-engineering-image-analysis-0cf491a17603
[24] E-SPIN, "Prompt Engineering Mastering: Techniques, Real-World Applications, and Future Trends in AI-Driven Interactions," *E-SPIN Group*, Sep. 12, 2024. Accessed: Mar. 29, 2025. Available: https://www.e-spincorp.com/prompt-engineering-techniques-applications-future-trends/