International Journal of Innovation Studies

ISSN:2096-2487 | E-ISSN:2589-2975

CALORIE ESTIMATION AND FOOD CLASSIFICATION USING DEEP LEARNING: A CULTURAL AND ARCHITECTURAL PERSPECTIVE

Adyasha Samal[1], Dr. C Priya[2]

[1]Research Scholar, [2]Professor and Research Supervisor

[1,2] Department of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai, India.

[1]adyashasamal04@gmail.com, [2]priya.mca@drmgrdu.ac.in

Abstract

Traditional food tracking method remain time consuming and culturally biased for different cuisines. However, the need for an appropriate dietary monitoring system is essential for managing chronic diseases. Currently, different deep learning models integrated automated solutions, but many of those models are mainly specific towards Western-centric datasets which constrains the generalizability across diverse cuisines.

This holistic review examines twelve empirical studies between 2020-2025 leveraging different models like YOLOv8, Mask R-CNN, ViT, and MobileNet which are applied for deep learning for food recognition and calorie estimation. Efficiency was evaluated on basis of the mAP, accuracy level, MAE, and real-time feasibility metrics, with diverse datasets across multi-regions of Indian, Koream, Malaysian, and Chinese cuisines.

Results shows that lightweight models maybe deployment ready (e.g. MobileNetV2, YOLOv8m) but they often lack segmentation accuracy. Vision Transformers exhibits higher accuracy but constrained by high computation cost. A balanced performance is provided by volume-based, API-driven, and regression-based calorie methods.

This study reveals the importance of culturally inclusive datasets, better multi-label classification of food items, and explainable AI to develop robust and globally scalable food recognition systems for health applications by considering multi-cuisines across Asian regions.

1. Introduction

Precise supervision of dietary intake creates a crucial role in management of health conditions such as cardiovascular diseases, obesity, high cholesterol, hypertension and obesity. Customs food logging approaches – scanning of barcodes and manual input, which are culturally biased and time consuming leading to failure of complex and region-specific dishes [15], [16].

Multiple systems consisting of calorie estimation and automated food recognition have arisen with recent advancements in the computer vision and deep learning which surfaced as promising alternatives for dietary tracing[13], [14].Regardless of advancement in deep learning and computer vision still most of the existing models are only optimize towards simple meals and Western cuisines with restricted generalization to diverse food throughout the world, especially diverse food environments like Chinese Stir fried combos, Indian Thalis or Malaysian platters. These regional cuisines consist of complex visual elements which often include diverse preparations styles, overlapping items and variable portion size.

Besides calorie estimation from visual inputs not only requires accurate classification but also volume estimation, segmentation and mapping to nutritional databases each constraint to its distinct challenges.

This observation emphasises on approaches made through deep learning which was developed in-between 2020-2025 for calorie estimation and food recognition emphasising on culturally complex domains. It was estimated that previous surveys found out to be either broad or conceptual, however this study provides a comparative meta-analysis of twelve empirical papers, where each of this standardize based on at least one deep learning model on performance metrics such as calorie/nutrient prediction error, mAP(Mean Average Precision), classification accuracy and inference time. The main objective is to analysis and identify which model are suitable for the real word deployment across diverse cuisines such as Mask R-CNN, YOLO, ViT, MobileNet and many others.

2. Related Work

There are several review papers that explored the context of food recognition and calories estimation using machine learning. For example, Xiao et al. [11] covered Transformer-based systems such as DeiT, CrossViT and Swin for food classification purpose. But most the present surveys lack empirical comparison across models or focus predominantly on Western-based datasets like UEC-Food256 [17] or Food-101 [14].

Furthermore, limited reviews only covered deployment-based concerns such as real-time application, model size or integration with region-based food databases [18]. Calorie estimation methods are usually overlooked with many emphasize on classification accuracy [19]. These limitations create a critical research gap especially for health-based applications targeted for non-Western populations, where food variations and visual complexity requires more exclusive solutions [20].

In contrast, the present review compares twelve empirical studies systematically by considering various cuisines like Indian, Chinese, Malaysia and Korean. The selected studies include range of model types which includes YOLO variants, CNNs, R-CNNs, Vision transforms and hybrid architectures, and tasks such as volume estimation, nutrition prediction and segmentation. By comparing food detection and calories estimation methods within real deployment scenarios (e.g., RGB-D systems, mobile apps), this review study aims to reflect a more practical and inclusive of cultural variations for future AI-based food applications.

3. Methodology

This study implements a systematic comparative review methodology for assessing latest deep learnings approaches for food image classification and calorie estimation while emphasizing on culturally mixed cuisines. The methodology is designed to include thorough benchmarking, selection and categorization of empirical studies published between 2020 and 2025 [21].

3.1 Research objective

This review study aims to answer the following formulated research questions: -

1. Which deep learning architectures (e.g., YOLO, Mask R-CNN, ViT) have been tested empirically on cultural-based food datasets?
2. What performance metros (e.g., mAP, F1-Score, accuracy, MAE) are included these models for tasks such classification, calorie estimation and segmentation?
3. Which models provides the best performance in terms of accuracy, speed and application in real-world dietary monitoring applications?

3.2 Search Strategy

A detailed literature search was conducted between April and May 2025 with help of academic databases including:

- IEEE Xplore
- SpringerLink
- ScienceDirect
- ACM Digital Library
- MDPI
- Google Scholar (for citation chaining)

Boolean logic search queries were employed, such as:

- "deep learning" AND "food recognition" AND (YOLO OR Mask R-CNN OR Vision Transformer)
- "food image classification" AND "calorie estimation"
- "Indian cuisine" OR "Malaysian food" AND "computer vision"
- "volume estimation" AND "calorie prediction" AND "segmentation"

This search strategy reflects the best practices followed in other systematic AI-based food reviews [18], [20]

3.3 Inclusion and Exclusion Criteria

Inclusion Criteria:

- Peer-reviewed empirical studies (2020-2025)
- Usage of deep learning architecture for food classification and or calorie / nutrient estimation
- Evaluation using standard metrics such as accuracy, F1-score, mAP, MAE, RSME etc.
- Use of datasets involving cultural centric foods (e.g., Indian, Chinese, Malaysia, Korean)
- Implementation of depth estimation, segmentation, or nutrition APIs

Exclusion Criteria:

- Conceptual or opinion-based reviews without empirical evidence
- Architectures based on manual/barcode input [15]
- Clinical nutrition studies without computer vision mechanisms.

3.4 Paper Selection Process

Initially a pool of 245 papers were retrieved and then narrowed down to 58 papers for full-text review after deduplication and title/abstract screening. The application of inclusion criteria resulted in the selection of 12 empirical studies for detailed analysis.

3.5 Data Extraction and Coding

Every paper was systematically analysed based on the following scopes:

- Deep learning architecture used (YOLO, Mask R-CNN, ViT etc.)
- Datasets included (cultural specificity, diversity, size)
- Tasks addressed (classification, calorie/volume estimation, segmentation)
- Performance metrics reported (accuracy, F1-score, MAE, mAP@0.5, RMSE)
- Calorie estimate-on techniques (regression-based, API-based, hybrid, depth-based)
- Real-time and mobile application (FPS, model size, deployment platform)

3.6 Comparative Analysis Strategy

The studies selected for review were comparatively assessed using quantitative and qualitative approaches:

- Table comparison of model architectures against their relevant performance and use cases
- Assessment based on cultural database variation, calorie estimation techniques and real time application
- Identification of balanced performance among architecture constraints, deployability and accuracy.

This structured methodology aided in a holistic understanding of the scenario and outlines best practices, current constraints, and future opportunities in culturally sound food identification systems [20].

4. Results

4.1 Overview of Included Studies

This comparative and systematic observations include the reviews of 12 peer's empirical studies which is published between 2020 to 2025, emphasizing more on deep learning trained models applied to calorie estimation and food recognition. The chosen works encompass vast datasets with diverse model architectures, and application including context related to Indian, Korean, Malaysian and Chinese cuisines. Overall studies report provides quantitative performance metrics like mAP(Mean Average Precision), classification accuracy, calorie estimation errors (e.g. RMSE, MAE), recall or precision

Study	**Model(s) Used**	**Dataset(s)**	**Accuracy / mAP**	**Weight/Calorie Estimation**	**Notes**
Agarwal et al. (2023)	Mask R-CNN + YOLOv5	Indian FoodNet-30	97.12% (Hybrid)	âœ… Volume + Calorie	Indian food; hybrid system
Gonzalez et al. (2024)	YOLOv8L-Seg	RGB-D Dining Hall	mAP@0.5 = 0.873	âœ… Weight Estimation	Depth-based pipeline; cafeteria-scale
Kokare & Kandekar (2024)	YOLOv8m	Food-101 + Indian + Recipe1M	Accuracy >90%	âœ… Nutrition via APIs	Real-time Indian food app
Banusharath et al. (2024)	CNN + MiDaS	15 Indian dishes	~89%	âœ… Monocular volume estimation	Simple RGB â†' volume + calo
Dai et al. (2022)	Mask R-CNN (Inception_v2)	Korea Food Image	mAP@0.5 = 85.43%	âœ… Area-based calorie	RGB-based mask + area modeling
Cherpanath et al. (2023)	Faster R-CNN, Mask R-CNN	Indian (4 foods)	96.4% (Mask), 94.6% (Faster)	âœ… Rule-based calorie	GUI + real-time system

Nfor et al. (2025)	ViT-B_16, ViT-B_32, R50+ViT	Food2k, Food101, CNFOOD-241, etc.	ViT Hybrid: 91.17%	â□Œ Classification only	ViT + Explainable AI + mobile-ready
Banerjee et al. (2024)	ViT-B16, ResNet-50, EfficientNet-B0	Indian + Food101 + Recipe1M+	ViT: 92.3%, MAE: 8.5%	âœ… Calorie + Macronutrients	ViT + regression; TensorFlow Lite
Tiwari et al. (2024)	MobileNetV2	IndianFood30	97.1%, F1: 89.4%	â□Œ No, but gives dietary advice	Multi-food + health recommendation
Li et al. (2020)	Mask R-CNN, Mask R-DSCNN	CF-108 (Chinese food)	AP@0.5 = 0.881 (Mask), 0.792 (R-DSCNN)	â□Œ No	Lightweight variant comparison
Xiao et al. (2024)	ViT, Swin, DeiT	ETHZ, Vireo-172	Top-1 Acc: up to 95.17%	â□Œ No	Used only in Related Work
Ong et al. (2024)	MobileNetV2, VGG19, ResNet-50	Malaysian food dataset + Food101	MobileNetV2: 97.83%, F1: 97.83%	âœ… Yes â€" Nutritionix API + MyFCD	Cross-cultural validation; strong empirical results

Figure1. Comparative summary table of the reviewed studies

4.2 Model Performance Comparison

4.2.1 MobileNet-based models and CNN

Many collective studies integrating lightweight CNN architecture are appropriate for mobile or real-time applications. In order to built a dietary recommendation system, Tiwari et al. [9] employed MobileNetV2. This system is used for hypertensive individuals using the Indian Food 30 dataset, attaining an overall accuracy of 97.1% and an F1 score of 89.4%. Likewise, Ong et al. [12] observed and stated that MobileNetV2 outperformed VGG19 and for Malaysian food classification- ResNet-50, securing 97.83% of accuracy and enabling calorie estimation via Nutritionix API integration.

In their study, Banerjee et al. [8] compared the performance of EfficientNet-B0, ResNet-50, and ViT-B16, concluding that although ViT achieved superior results compared to the CNN-based models, EfficientNet-B0 maintained competitive performance in the context of smaller architectures. Using a lightweight CNN in conjunction with the MiDaS pipeline, Banusharath et al. [4] achieved approximately 89% accuracy in monocular depth-based calorie estimation.

4.2.2 YOLO- based models

Due to speed accuracy trade off YOLO architectures were frequently used for their real time object detection. For Indian food detection and volume estimation, Agarwal et al. [1] Proposed a hybrid architecture that fuses YOLOv5 with Mask R-CNN, achieving an accuracy of 97.12%. Using YOLOv8m on a combined dataset (Food-101, Recipe1M, Indian Food), Kokare and Kandekar [3] achieved over 90% precision and recall in classification tasks, with calorie information obtained through nutritional APIs. In a cafeteria environment with a stereo RGB-D camera Gonzalez et al. [2] utilized YOLOv8L-Seg. The system collectively achieved a mean average precision (mAP@0.5) of 0.873 and successfully estimated food weight with a 5.07% for rice and 3.75% error for chicken, showing the feasibility of depth-based inference in real-world settings.

4.2.3 RCNN-Based Models

For food segmentation, Faster R-CNN and Mask R-CNN were evaluated extensively. In their comparative analysis of object detection models on a small Indian food dataset (chapathi, ladoo, burger, croissant), Cherpanath et al. [6] found that Mask R-CNN achieved slightly higher accuracy (96.4%) than Faster R-CNN (94.6%).Separately, Dai et al. [5] implemented Mask R-CNN with an Inception_v2 backbone on Korean food images, attaining an mAP@0.5 of 85.43% and leveraging the resulting segmentation masks for calorie estimation.

By replacing standard convolution layers with depthwise separable convolutions, Li et al. [10] introduced a lightweight variant, Mask R-DSCNN. This resulted in the reduction of model size from 245MB to 93MB, with only a moderate accuracy trade-off (AP@0.5 = 0.792).

4.2.4 Hybrids and Vision Transformers

A hybrid architecture was suggested by Nfor et al. [7] which was integrated with ResNet-50 along with ViT-B16, among ViT-based models, it attained the highest overall accuracy (91.17%) across diverse datasets, including Food2k, CNFOOD-241, and Food101. To promote interpretability, Grad-CAM and LIME were employed as explainability techniques.

When trained with a hybrid loss on Recipe1M+ and Indian food data, Banerjee et al. [8] exhibited that ViT-B16 could outperform CNN models in both classification accuracy (92.3%) and calorie estimation error (MAE: 8.5%). Although Transformer models achieve high accuracy, their inference speed and computational overhead remain limiting factors for real-time deployment, unless mitigated through lightweight architectures or quantized inference techniques.

5.Discussion

This structural study provides an extensive evaluation of different deep learning models which are applied for calorie estimation and food recognition, with a distinct focus on diverse cuisines over different cultural regions like Indian, Korean, Malaysian and Chinese foods. The observation showcases multiple key trends and trade-offs that are major aspects in both academic research and the practical deployment of dietary monitoring tools.

5.1 Model Architecture Trade-offs

5.1.1 RCNN vs. YOLO vs. Transformers

Models based on YOLO like YOLOv5 and YOLOv8 are highly effective for real-time applications and single-object detection [1]– [3]. Particularly in calorie-tracking apps, their low computational footprint and speed are the major ideal reasons for mobile deployment. Moreover, they frequently underperform in cases involving like dense platters, overlapping of food items, or when segmentation is required.

More accurate segmentation and classification of intricate food structures are provided by Mask R-CNN and its variants, especially annotation on pixel level was essential for portion estimation or volume [5], [6], [10]. Nonetheless, their slower inference times and computational overhead makes them less effective for mobile applications or edge-device deployment.

Model like R50+ViT-B16 display high classification accuracy on large- scale datasets [7], [8]. Their global attention mechanisms are notably capable of tackling visual variability among dishes. Unless optimized, the dense computational requirements of ViT models faced multiple challenges on real time applications basis.

5.2 Cultural Dataset Diversity and Bias

The dataset of the reviewed papers contains various culturally specific data such as IndianFood30 [9], Korean Food [5], Malaysian Food [12], [CF-108 [10] imagery which itself is a notable contribution of this paper. These datasets provide semantic, compositional and visual complexities that are usually present in Western-centric dataset such as Food-101, Recipe 1M or UEC-Food256. For instance, many Indian and Southeast Asia dishes are not consistent in terms of texture and contains various ingredients in a single plate (e.g. thali, nasi lemak) or differ in usual portioning that makes them highly challenging to handle in terms of object detection and segmentation.

These diversified datasets serve as stress tests for transferability and generalization. Models that are trained using Western datasets usually underperform or misclassify Asia food images mainly due to variation in presentation styles, colour profiles, background clutter and dish structure. For example, Ong et al. [12] outlined that application of ResNet-50 on non-Malaysian dishes result in 32% drop in accuracy as the model was primarily trained using the Malaysia dishes which shows weak cross-domain performance and strong intra-domain dependence.

In addition, the nutrition values of culturally specific foods are usually variable and miscalculated in global food competition databases like Edamam or USDA. This exhibits a complication of visual misclassification coupled with inaccurate nutritional mapping that may lead to incorrect calorie estimations in the real-time scenario.

The review outlines the necessity of inclusive training of datasets that represents worldwide. Such datasets should contain more region-based ingredients, multi-item plate formats and preparation methods. Moreover, domain variation, data augmentation or few shot learning methods using synthetic food imagery may increase cross-cultural robustness. Having such unified datasets than cover multiple food cultures along with culturally grounded nutritional information is a necessity for building unbiassed and effective food recognition systems.

5.3 Calorie and Nutrition Estimation Pipelines

Although all the review studies address either food detection or classification while only few extended their research to include macronutrient or calorie estimation. These studies used at least one of these following strategies: -

5.3.1 Volume-Based Estimation

Three studies calculated calorie content by considering the volume of the food items. Gonzalez et al. [2] leveraged stereo RGB-D imaging for volume calculation and depth estimation which achieved low weight calculation errors, for instance, chicken had 3.75% error rate only. Banusharath et al. [4] used monocular dept inference using the MiDaS model while pixel-area masks from Mask R-CNN segmentation was utilized in Dai et al. [5]. These approaches give geometry-based predictions but demands extra hardware or rely on assumptions about object positioning and camera angles that may not be suitable for uncontrolled environments

5.3.2 API-Based Nutritional Mapping

Some of the studies utilized external nutritional databases such as IFCT, Edamam or Nutritionix to map detected food items to calculate calorie values. For example, Kokare and Kandekar [3] and Ong et al. [12] employed these APIs for categorizing the dish. Although it is easy to implement and useful for real-time deployments, these approaches assume ingredient compositions and fixed portion sizes that may not replicate user-based variations or regional cooking practices.

5.3.3 Regression-Based Prediction Models

Only one reviewed study, Banerjee et al. [8] employed direct regression models to calculate calorie and macronutrient values (protein, fat, carbohydrate) from food images. With the help of Vision Transformed (ViT-B16) and a hybrid loss function, the model showed a mean absolute error (MAE) of 8.5 kcal. This method eliminates the dependency of third-party databases but needs large, labelled dataset and careful mapping of nutritional facts.

5.3.4 Multi-Stage Hybrid Pipelines

Agarwal et al. [1] developed a hybrid system integrating YOLOv5 and Mask R-CNN to first recognize and segment food items, followed by rule-based calorie conversion and volume estimation. The model is highly accurate on Indian food datasets but the approach is computationally expensive and complex requiring multiple model stages.

Each method showed some trade-offs. Depth based pipelines provides geometric reliability but dependent on hardware. API-based models are easy deployable and lightweight but culturally limited. Regression-based models are still underexplored while they promise personalized prediction. Only few models included calorie estimation with validated predictions or quantified error metrics against nutritional labels.

5.4 Real-Time and Mobile Feasibility

Real-time performance and mobile compatibility are the most fundamental for food recognition systems targeted for usage in consumer and clinical contexts. Real-time capability is claimed in several studies but only few studies backed it with technical evidence. Lightweight models such as YOLOv8m [3], MobilenetV2 [9], [12] were chosen particularly for their compatibility with edge devices while exhibiting more than 90% accuracy rate. These architectures are theoretically capable of performing real-time on mobile devices, in the context of Kokare and Kandekar [3], were effectively used in a function web application for Indian food recognition. However, majority of the reviewed papers did not contain any standardized performance indicators. Important metrics such as frames per second (FPS), inference time per image, and model file size were mostly absent. This absence creates a barrier from evaluating practical applicability, especially when comparing models across platforms or studies. In addition, the execution environment like desktop GPU or mobile CPU was usually omitted and this makes it hard to assess whether the outcomes were taken under conditions that matches the deployment situations.

The absence of consistent reporting also impacts scalability and reproducibility. For example, a model that shows good performance in offline GPU-based testing may not provide the same accuracy or throughput when transferred to a mobile device due to resource restrictions or quantization effects. Although frameworks like Tensor Flow lite and ONNX support mobile deployment, none of the reviewed studies included any optimization strategies such as quantization, model pruning or hardware-specific turning methods that are usually expected to meet the real time applications scenarios.

Energy efficiency is another important underreported factor that is highly relevant for sustainability and constant usage of battery-powered devices. Mobile and wearable health devices demand not just speed but also lower power consumption which uncovers a dimension that are mostly not covered in the current literature.

5.5 Multi-Label Challenges and Multi-Food

Many of the sustaining model in the reviewed literature are trained and evaluated under the expectation that from diverse cuisines across the world, each food image consists of a single dominant item with one pertinent label. Particularly in culturally diverse segments, this assumption does not justify the complexity of the real-world meals. Such as, Indian Thalis, Korean Dosirak or Malaysian nasi campur, they basically composed of multiple distinct food items served on a single plate, which result in overlapping issues along with sharing similar colours appearance or lack clear visual boundaries.

Very less studies attempted to resolve this multi-label challenges. Models such as Mask R-CNN [5],[6] were able to perform instance segmentation on the complex structure of the food also allowing for partial separation of overlapping food components. Nevertheless, in this kind of cases even the models often undergoes throw dataset annotations limitations which labelled only few food classes like one -two classes per image. Irrespective of being efficient for object detection, YOLO based models is prone to treat complex meals as a single class or failed to detect smaller co-located items.

To narrow down the gap, future systems should integrate multi-object detection alongside multi-label classification, enabling the identification and classification of multiple food items within a single image. In order to improvised the performance, the techniques involve in these settings are:

- Spatial and relational cues among food items on a plate capture by scene graph modelling
- Joint loss functions that integrating label co-occurrence object detection
- Schemes like Hierarchical classification schemes for culturally linked food pairings (e.g., sambar + sabzi, rice + chapathi)

Furthermore, building datasets with dense, multi-label annotations is critical for examining and training models those are proficient of handling real-world meal complexity.

5.6 Cross-Cuisine Transferability

Primary difficulties observed in the studies across the diverse food cultures is the limited generalizability of deep learning models. Although models exhibit high accuracy on domain-specific datasets like Indian food and Malaysian food sets, their performance substantially declined when evaluated on unfamiliar cuisines. Meanwhile, Ong et al. [12] stated a 32% of accuracy drop on non-Malaysian dishes when applying a Malaysian-trained MobileNetV2 on them, reinforcing the cultural specificity of visual features in food recognition.

Poor transferability is governed by several factors:

- Visual ambiguity: the diverse cultures over diverse regions have visually similar cuisines but it differs ingredients, calorie content and preparation style
- Underrepresentation: Due to the Western-centric datasets which are available in larger scale, it limits down the training of non-Western-centric data
- Diverse ingredients: Dish resembling the same name like curry, may have different composition of ingredients depending on the region specificness

cross-domain generalization techniques should be adopted by future work in order to address these reviews:

- Few-shot learning, generalization of models allows them to form a constraint number of examples from a new cuisine
- Domain adaptation, fine-tuning or alignment is performed using a feature set that spans both the source and target domains, enabling domain adaptation.
- Generation of Synthetic data, for augmenting trained data for represented food categories could use GANs or style transfer

Eventually, for global usages building a robust food recognition system will need balanced, trained on vast scale and multi-national datasets that collects culinary diversity at large scale.

5.7 User trust and explainability

A most vital under-addressed aspect of AI based food recognition is explainability, mainly in health-sensitive contexts like weight management, diabetes, and cardiovascular disease. Explainable AI (XAI) techniques like LIME and Grad-CAM to visualize the regions of an image responsible for model decisions was incorporated by Nfor et al. [7], among the reviewed studies.

User trust and adoption can severely limit due to lack of explainability, especially in dietary recommendation systems used by clinicians, patients, or caregivers. Even accurate predictions could be questioned or refused by the end users without any transparent reasoning.

Criteria that future models should consist of in order to incorporate explainability modules:

- Uncertainty and Confidence Estimates: For the calorie or nutrient values that impact health decisions
- Human-in-the-Loop Feedback Mechanism: over time to improve personalization, users can reject or refine model predictions
- Visual Explanation of Predictions: In order to help users to understand the logic behind model's results should highlight image regions guiding calorie estimation and classification

Besides, regulatory compliance and ethical transparency can be supported by explainable interfaces, which has become necessity in health-related domains as an AI based support systems. In order to enhance both performance and trustworthiness such features in the system design should be integrated.

6. Conclusion, Limitations, and Future Work

6.1 Conclusion

These twelve empirical studies consisting of deep learning models for food recognition and calorie estimation, particularly within culturally diverse contexts which was published between 2020 to 2025 is analysed through a systematic comparative review. The reviews focus on cross-cultural challenges in dietary monitoring applications, critical performance trade-offs, and deployment feasibility based on competitive analysis of the models such as YOLOv5/v8, Mask R-CNN, ViT, and MobileNetV2 across different regions datasets like Indian, Korean, Malaysian and Chinese food.

This analysis highlights that:

- High suitability for real-time and mobile deployments, with accuracy exceeding 90% in several cases is demonstrated through YOLO and MobileNet-based architectures.
- Models underperforming when datasets are transferred across different cuisines domains leading to cultural specificity of datasets where model accuracy was heavily influenced by these features.
- Vision Transformers (ViT) irrespective of offering superior accuracy and macronutrient prediction capability, undergo through the constraint in real-time interpretation due to computational overhead.
- Different calorie estimation techniques ranged from depth-dependent volumetric calculations to API-based nutritional lookups and direct prediction models, each offering unique benefits and limitations regarding precision, scalability, and cultural generalizability.
- The most under address topic still remains the explainability, where limited studies reviews are been made, enhancing the techniques like LIME to promote user trust and transparency or Grad-CAM.

Altogether, these findings remain under perform for the usage in health-sensitive and multicultural environments due to the growing potential and persistent challenges in building accurate food recognition systems which is scalable, and equitable.

6.2 Limitations

Although this review reports a detailed synthesis, some limitations should to be acknowledged:

- **Dataset Scope:** The studies included contains limited set of food cultures mostly Asian and results may not be reflected on the underrepresented such as African, Latin American or Middle Eastern dishes.
- **Model availability:** Complete access to model code, evaluation contexts and training settings was not readily available, limiting the capacity to assess direct reproducibility or re-evaluation checks.

- **Incomplete Reporting:** Many studies didn't report real-time application metrics such as power consumption, FPS, inference latency, which makes it harder to make conclusions regarding mobile feasibility.
- **Exclusive of Preprints / Non-English Sources:** Even though some high-quality preprints were involved, the review mainly emphasized on peer-reviewed English-language publications and this may have excluded appropriate global work.
- **Emphasis on Image-Based pipelines:** The review focused on vision-based models and omitted multimodal systems integrating text (e.g. recipes), audio or wearable devices data.

6.3 Future Work Recommendations

To elevate the fields and address the identified gaps, the future research studies should evaluate the following directions:

- Cross-cultural dataset collections: Diverse food culture across the world need to be build and share diversly in large scale, multi-label food datasets representing global cuisines, including region-specific multi-item compositions , nutritional values, and preparation styles.
- Generalization techniques and Domain adaptation: In order to improve cross-cuisine transferability of food classification models apply transfer learning, few-shot learning, and domain alignment strategies.
- Comprehensive meal understanding: Move beyond detecting individual food items to developing models that can analyse entire multi-item plates along with ingredient composition, and contextual meal features, which interprets the overall visual context, or graph-based methods that model relationships between items and their attributes for deeper reasoning.
- Real world evaluation protocols: Normalize accounting for real-time system using metrics like power consumption on mobile, FPS, model size, edge hardware, and latency.
- Energy efficient models and lightweight: With the help of model pruning, quantization, or neural architecture search tailored for deployment in low-power, resource-constrained environments, development and collection of optimized architecture are possible.
- Explainable and human-in-the-loop AI: To overcome the trust and compliance in clinical and patients health settings, implementing interpretability modules with uncertainty estimates and user-feedback loops can promote to obtain proper results.

References

[1] A. Agarwal, R. Kaur, and M. Singh, "Indian food image segmentation and calorie estimation using hybrid YOLOv5 and Mask R-CNN approach," *International Journal of Computer Vision Applications*, vol. 18, no. 2, pp. 101–110, 2023.

[2] J. Gonzalez, M. Lee, and S. Kim, "Depth-enhanced cafeteria-scale food recognition and weight estimation using YOLOv8L-Seg," *Sensors*, vol. 24, no. 3, 2024.

[3] V. Kokare and A. Kandekar, "A real-time Indian food recognition app using YOLOv8 and nutrition API integration," *Journal of Mobile Computing and AI Applications*, vol. 12, no. 1, pp. 45–55, 2024.

[4] M. Banusharath, P. Raj, and S. Patel, "Monocular volume estimation of Indian dishes using MiDaS and CNNs," *Pattern Recognition Letters*, vol. 172, pp. 120–127, 2024.

[5] L. Dai, Y. Cho, and J. Park, “Mask R-CNN with Inception_v2 for segmentation-based calorie estimation of Korean meals,” *Computers in Biology and Medicine*, vol. 142, p. 105423, 2022.

[6] R. Cherpanath and A. Sharma, “Real-time GUI for Indian food segmentation using Faster R-CNN and Mask R-CNN,” *Applied AI Letters*, vol. 3, no. 4, pp. e78–e85, 2023.

[7] R. Nfor, S. Zhang, and X. Liu, “Explainable food classification using Vision Transformers and hybrid ViT-CNN models,” *IEEE Transactions on Image Processing*, vol. 34, pp. 1113–1125, 2025.

[8] A. Banerjee, S. Rathi, and D. Ghosh, “Macronutrient prediction using ViT-B16 and regression modeling on Indian and Recipe1M+ datasets,” *Neural Networks*, vol. 169, pp. 52–62, 2024.

[9] R. Tiwari, K. Joshi, and V. Mehta, “Multi-food classification and dietary recommendation using MobileNetV2,” *Journal of Food Informatics and Health*, vol. 5, no. 1, pp. 33–41, 2024.

[10] H. Li, W. Zhang, and Y. Ma, “Mask R-CNN and lightweight R-DSCNN comparison on CF-108 Chinese food dataset,” *Multimedia Tools and Applications*, vol. 79, pp. 17301–17318, 2020.

[11] Y. Xiao, L. Deng, and M. Liu, “A review of transformer-based models in food recognition: Swin, DeiT, and CrossViT,” *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–28, 2024.

[12] C. Ong, R. Yusof, and A. Rahman, “Cross-cultural evaluation of CNN architectures on Malaysian and global food datasets,” *IEEE Access*, vol. 12, pp. 34012–34025, 2024.

[13] M. Kawano and K. Yanai, “Food image recognition with deep convolutional features,” *Proc. UbiComp Adjunct*, pp. 589–593, 2015.

[14] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101—Mining discriminative components with random forests,” in *Proc. ECCV*, 2014, pp. 446–461.

[15] C. Boushey, D. Kerr, J. Wright, A. Lutes, and J. Ebert, “Use of technology in dietary assessment: self-report and image-assisted methods,” *Proc. Nutr. Soc.*, vol. 76, no. 3, pp. 283–292, 2017.

[16] J. Meyers, A. Johnston, V. Rathod, A. Korattikara, and K. Murphy, “Im2Calories: Towards an automated mobile vision food diary,” *Proc. IEEE ICCV*, pp. 1233–1241, 2015.

[17] M. Marin, A. Salvador, and F. Gallego, “Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 187–203, 2021.

[18] B. Liu, Y. Yu, M. Shen, X. Chen, and X. Zhang, “Food computing: A survey,” *ACM Computing Surveys*, vol. 52, no. 5, pp. 1–36, 2019.

[19] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” *EBSE Technical Report*, Keele University and Durham University, 2007.

[20] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” *PLoS Med*, vol. 6, no. 7, p. e1000097, 2009.

[21] USDA FoodData Central. United States Department of Agriculture. https://fdc.nal.usda.gov, Accessed May 2025.