



RESEARCH ARTICLE

OPEN ACCESS

VIT-FOODNA: AN END-TO-END TRANSFORMER-BASED MULTIMODAL FRAMEWORK FOR FOOD RECOGNITION AND NUTRITION ANALYSIS

E. Anitha ^{*1}, and Dr. A. Bazila Banu ²

¹Department of Artificial Intelligence and Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India

²Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamilnadu, India.

¹<https://orcid.org/0000-0002-0092-9106>^{ID}, ²<https://orcid.org/0000-0003-2647-6667>^{ID}

E-mail: *anithaeswaran1803@gmail.com, bazilamakki@gmail.com

ARTICLE INFO

Article History

Received on: March 04, 2026

Reviewed on: April 02, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

Keywords:

Multimodal Food Analysis,
Dish Classification,
Nutrition Profiling,
Cross-Modal Feature Fusion,
End-to-End Dietary Analysis.

ABSTRACT

Dietary monitoring and health care will both require accurate food recognition and nutritional analysis. The current methods mainly involve convolutional networks that analyse images and word embedding that process ingredients as distinct processes with dish identification, portion size, and nutrition labelling as independent. This paper presents a multimodal transformer-based framework called as ViT-FoodNA that integrates these tasks into one end-to-end framework. It suggests using Vision Transformer (ViT), which is used to encode the visual attributes of food images, and Transformer based encoder to learn semantic relationships between inputs of ingredients. A cross-modal fusion transformer consists of visual and textual embedding to facilitate contextual interaction between the images of foods and their ingredients. Based on the result of the fused representation, the model collectively predicts the identity of the dish along with portion size and comprehensive nutritional breakdown using a shared transformer decoding strategy. In contrast to the previous convolutional neural network (CNN)-based embedding structures, the suggested architecture uses the self-attention mechanism to learn fine-grained intermodal dependencies, which enhances its resistance to visual ambiguity and missing ingredients in a list. Experiments indicate that ViT-FoodNA produces up to 12% higher Precision, Recall, F1-score, and 9.75% and 5% greater Top-1 and Top-5 accuracy and up to 32.7% and 30.4% lower MAE and RMSE than existing models. The reduction of PMAE is 6.5%, and the increase of mAP is 9.8%, which proves the great benefit of transformer-based multimodal fusion in the full analysis of diet.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Food analysis has emerged as an urgent field of concern in the health and sustainability of the environment in the face of growing prevalence of lifestyle diseases, diets, and individualized nutrition around the globe. The growing number of digital images of food on smartphones and social media has served as an incentive to develop automated dietary analysis tools that can identify nutritional information about daily food intake. These systems may be useful in the nutrition monitoring and prevention of chronic diseases and conducting dietary studies with a large population, in particular in those areas where manual dietary logs continue to be error-prone and time-consuming [1]. With the diversification in the food consumption patterns brought on by urbanization, globalization and the advent of the convenience foods, the use of nutritional tracking has become even more complex. The aesthetic variety of meals, which fluctuates in how it is cooked or prepared, cooking style, and quantity of portions and cultural components, makes manual calorie calculation very inaccurate. Research has demonstrated that people tend to overstate their nutritional intake by 20-50 percent that accumulates to nutritional deficit or surplus. This further increases the necessity to have smart, automated methods that will help people, clinicians, and researchers have a better grasp of dietary habits [2]. Deep learning (DL), in particular, CNNs, has become an influential part of the development of food recognition systems. CNNs have been proven to be effective in the classification of dishes and identification of ingredients when trained on large-scale datasets.

Nonetheless, food images are highly intra-class (e.g., a rice dish with a variety of toppings) and low inter-class separable (e.g., visually similar curries or stews), and hence the generalization ability of purely CNN-based models. Besides, CNNs are weak at capturing long-range contextual information - e.g., whether two or more ingredients are present together, or whether an individual has more than one food item on a plate - that is vital to the fine-grained understanding of dishes [3]. In parallel with the visual modeling, the ingredient lists and the descriptions of the recipes are highly informative in semantic, and they can enrich the knowledge of food. Nevertheless, more complex pipelines based on text processing have used word embeddings or shallow RNNs that do not assume interactions between ingredients, more elaborate recipe structures or subtle nutritional semantics. This loophole is an issue since ingredients are directly associated with nutrient profiles, cooking changes and caloric density. Models will not be able to know correct macronutrient and micronutrient distributions without a solid representation of textual semantics [4]. Another area of long-time difficulty in dietary assessment is that of portion size estimation. As compared to traditional approaches, which are based on either single view image geometry or reference objects, they are subject to scale ambiguities and occlusions. Equally, nutrition profiling pipelines remain largely reliant on lookup tables or manual mapping of the predicted ingredients to food composition databases and thus lead to compounding errors. Such multi-stage, modular systems are thus characterized by low reliability as well as inapplicability to the real world [5].

The most recent developments in DL have placed a strong emphasis on the potential of introducing a revolution in the field of Vision Transformers (ViTs) and text-based transformer models including BERT and GPT types. Transformers are good at capturing long-range and global interaction both in a visual and textual form. ViT-based models have made exceptional advancements in fine-grained recognition tasks, and transformer language models allow understanding details of complicated ingredient lists and nutrition metadata. These developments have motivated new research paths to utilize multi-head self-attention to consolidate and improve multimodal food analysis [6]. The concept of multimodal learning in the food domain has become popular, and new literature proves the usefulness of the combination of food images and ingredient texts or recipes descriptions. Combination models Visual and textual cues can result in high dish recognition, less ambiguity and better generalization across different cuisines. Nevertheless, these works are usually concerned with dishes classification only, but not with lower-level actions like portion estimation or prediction of nutrients. Moreover, they hardly use transformers as the main design to have unified multimodal fusion [7]. Notwithstanding these improvements, the majority of the current food-computing systems take a fragmented approach to their workflow, i.e. dish classification, portion estimation, and nutrition calculation modules are executed as independent tasks. This modularity results in error propagation wherein errors in the initial stages would have dire impact on downstream predictions.

Furthermore, such pipelines do not have profound cross-modal reasoning that is able to match visual information (e.g., color, texture, plating) to ingredient semantics and nutritional properties [8]. In recent years, multi-task learning frameworks have shown considerable potential in strengthening and improving food analysis systems. Sharing the learning of dishes classification, portion estimation, and nutrient prediction makes it possible to learn the representations jointly and decrease redundancy. Nonetheless, little exploration has been done on transformer-based multi-task architectures to conduct unified food analysis. Also, the majority of models fail to exploit the entirely end-to-end pipelines in which the visual, textual, and volumetric signals are trained on heterogeneous data streams. Lack of profound cross-modal alignment mechanisms is another significant weakness of modern food analysis systems. The images of foods and textual foods have complementary information and combining these two would require high-order attention mechanisms, which can learn inter-modal relationships. Transformer layers based on cross-attention can be used to represent a potent approach to aligning visual features with ingredient semantics, but there has been little adoption of these networks in food and nutrition analysis. It results in imperfect knowledge of the composition of dishes and incorrect estimation of nutrition [9]. What is more, existing systems have a hard time to include some contextual information, like cooking techniques, cultural differences, and ingredient changes, all of which have significant impacts on the dish appearance and nutritional content.

The food items available in most of the existing models are additional physical objects, but dynamic aspects such as interaction of the ingredients or the mode of preparation that can dramatically change the caloric and nutrient density are neglected. Traditional models cannot be used to generalize the effects of different cuisines and complicated dishes unless contextual metadata and hierarchical ingredient relations were incorporated into the learning process. This drives the necessity of multimodal structures that are driven by transformers and able to distil more contextual semantics of food so that more precise and holistic food knowledge is realized [10]. The study is aimed at addressing these issues by offering a solution referred to as FoodNA, a transformer based multimodal system that enables combined food and nutrition analysis. FoodNA is an image-based model that consists of a combination of Vision Transformer-based image with transformer-based ingredient and cross-modal fusion module to give a complete picture of food items. The model does classification of dishes, portioning and nutritional profiling simultaneously and does not have the errors experienced by the traditional multi-stage systems. With the assistance of attention-based multimodal learning, FoodNA is also able to match visual representations with the information about ingredients, which can help to estimate the portion size, calorie content, and predict nutrients more accurately. Following are the key contributions of proposed work:

- A complete end-to-end multimodal transformer architecture (FoodNA) that jointly learns dish classification, portion detection, and nutrition profiling.
- A Visual encoder using Vision Transformer, which represents the global spatial relationships and fine-grained food qualities to enhance recognition of dishes and understanding of portions.
- A Transformer-based ingredient and nutrition encoder that model's ingredient dependencies, cooking transformations, and nutritional relationships through contextual embeddings.
- A cross-modal fusion transformer that performs attention-driven alignment between visual and textual features, enhancing joint dietary understanding.
- A framework for multitask learning that can concurrently predicting dish labels, food volume, and macro-micronutrient composition with high accuracy.
- Superior performance and explainability, demonstrated using attention visualizations and comprehensive experiments against existing unimodal and modular baselines.

This work's remaining portions are arranged as follows. Section 2 analyses available literature and Section 3 expound on the proposed works. In section 4, the outcomes are explained and an estimate of the work's efficacy is provided. The work is concluded in Section 5.

II. LITERATURE SURVEY

A technique of automated image-based dietary assessment of Mediterranean food has been presented in [11] based on: 1) a collection of photos of Mediterranean foods; 2) a CNN trained to recognize food pictures; and 3) stereo vision techniques for figuring out a meal's volume and nutritional value. They were not providing unified end-to-end nutrition prediction and it is not multimodal (image + ingredient). [12] developed one RGB-D fusion network that comprised of multi-scale fusion and multimodal feature fusion (MMFF) in nutrition assessment based on vision. To produce effective feature fusion, MMFF used a convolutional block attention module and a balanced feature pyramid. The text/ingredient semantics is missing to monitor nutrition in detail and it is largely based on depth sensors. [13] suggested a hybrid architecture that forecasts the quantity of calories in different food items in a bowl using DL algorithms. Segmentation, categorisation, and figuring out the amount and calories of food items are the three main parts of this. The pictures are initially segmented when they employ a Mask RCNN. The food item will be categorized by collecting attributes on the segmented photos through the YOLO V5 framework. They restricted scalability and generalisation with a shallow hybrid architecture with no transformers and cross-modal reasoning. A Multi-Modal Alignment Method of Cross-Modal Recipe Retrieval (MMACMR) was proposed by [14]. The dissimilarity in the ambiguous food picture between the supervision of the relating recipes under this approach is conducted to consider both the intra-modal and inter-modal alignment.

They however do not emphasize on unified categorisation, portion control and nutrition prediction within an end-to-end system but only emphasise on recipe retrieval. [15] used deep CNNs with a self-attention mechanism to develop a reliable food recognition model (FRCNNSAM). They trained many FRCNNSAM structures with different settings, then averaged their predictions. Regularisation was used to ensure that the model did not get too complex such that it would over-fit and under-fit data to generate more training data. They just integrate attention on the visual level; they neither work with joint nutritional estimation nor ingredient modelling. To improve the efficacy and precision of food nutrient identification, [16] proposed a special integrated DL model. They used the merits of Feature Pyramid Network (FPN) in multi-scale fusion of features, Swin Transformer in long-range features relationships, and EfficientNet in feature extraction. The model omits sometimes ingredients of complicated dishes, needs lots of computational resources to train, and is less precise when it is dealing with low-quality or low-light pictures. Pic2Plate, the system presented by [17], is a system that integrates Retrieval-Augmented Generation (RAG) with Vision-Language Models (VLMs). Pic2Plate integrates RAG to gather and personalise recipe suggestions and high-tech image detection to remove ingredient lists of users in pictures. Pic2Plate's limitations include its reliance on an unbalanced FoodSeg103 dataset, its inability to handle a variety of real-world meal scenarios, its lowered accuracy for uncommon or region-specific items, and its dependency on internet access. [18] extinguished the Multi-level Attention Feature Fusion Network (MAF-Net).

In particular, they employ the feature maps produced at various levels by the CNN's backbone network as inputs. Nevertheless, its multi-step training strategy means that MAF-Net is costly to compute, and is not generalisable to non-assessed data. [19] presented a smartphone application that uses CNNs and other DL techniques to estimate calories from food photos. However, their reliability in real-world situations was limited by their difficulties with food presentation inconsistency, inaccurate portion size estimate, and sensitivity to changes in picture quality. [20] developed and implemented a DL system that is used to automatically classify and sort meals. The lack of predicted direct nutritional content, the moderately small dataset size, and a fact that the portion size of 2D photos can be calculated only in different circumstances are the limitations of the work. A novel multimodal AI-based system, which predicts food ingredients based on multiple input modalities, was proposed by [21]. The proposed approach is a two-step mechanism capable of properly detecting ordinary and outlier things because CNNs are applicable to detect visual features in pictures, and transformer-based models are suitable networks to detect features in textual inputs. They do not use categorisation, portion estimate and nutrition prediction as a combined end-to-end transformer pipeline but, in fact, they only predict ingredients. [22] proposed a nutritional advising system based on images which uses multimodal DL to automate the process of food recognition, capacity prediction and dietary recommendation. The model is highly accurate in respect to a wide range of food types such as unknown foods because it is powered by a highly refined CLIP model of recognizing foods without any prior testing. The propagation of errors between jobs was attributed to the fact that they used modular stages instead of using the single fusion transformer.

RESEARCH GAP:

Existing food analysis systems are still constrained by CNN-based visual models, shallow ingredient representations, and modular multi-stage pipelines that consider nutritional prediction, portion estimate, and dish categorisation as distinct tasks, leading to weak cross-modal reasoning and error propagation. In order to concurrently align visual cues with ingredient semantics, existing multimodal systems lack attention-based fusion and seldom employ unified transformer topologies. Additionally, they have trouble with fine-grained quantity estimating, missing ingredient lists, and visual ambiguity. To accomplish more precise and comprehensive dietary analysis, an end-to-end transformer-based multimodal framework that combines pictures, ingredients, and nutritional signals inside a single unified architecture is obviously needed.

III. PROPOSED METHODOLOGY

An End-to-End Transformer-Based Multimodal Framework for Food Recognition and Nutrition Analysis named as ViT-FoodNA were proposed here.

III.1 DATASET COLLECTION

An MM-Food-100K dataset [23] from Hugging Face was used in the analysis of the proposed ViT-FoodNA model. The dataset is a multimodal food intelligence collection consisting of 100,000 crowd-sourced food images along with detailed associated metadata. Each sample contains the dish name, food type, list of ingredients, portion size estimations, cooking method, and information about nutrition such as protein, fat, calories, and carbohydrate values. The images were gathered through a large community-driven campaign, and all annotations were generated using a hybrid process involving initial human inputs, automated vision-language model analysis, and final human validation. The dataset includes five major food categories—homemade food, restaurant food, packaged food, raw vegetables/fruits, and others—making it suitable for both food recognition and nutrition estimation tasks. For training and testing purposes in this work, the dataset was split into a 70:30 ratio.

III.2 PROPOSED END-TO-END TRANSFORMER-BASED MULTIMODAL FRAMEWORK (ViT-FOODNA)

The architecture for Proposed End-to-End Transformer-Based Multimodal Framework called as ViT-FoodNA for Food Recognition and Nutrition Analysis. It contains five major layers as follows:

- Input Layer,
- Feature Extraction Layer,
- Cross-Modal Fusion Transformer Layer,
- Multi-task Prediction Layer and
- Output Layer.

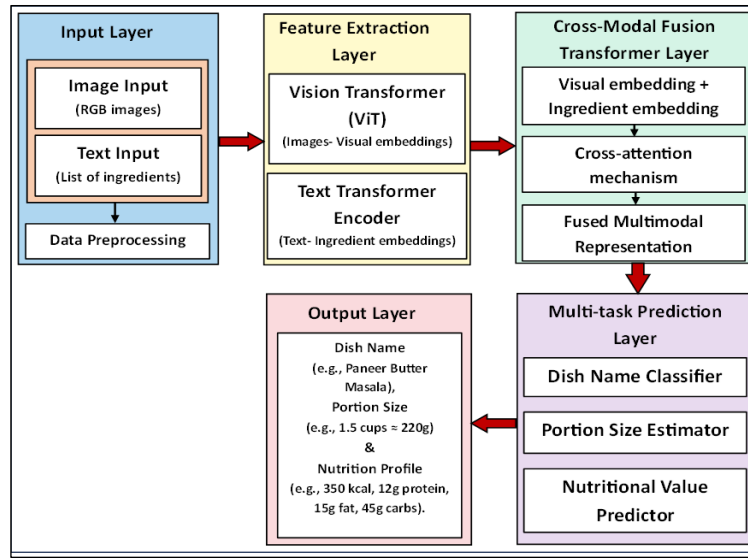


Figure 1: Architecture of proposed ViT-FoodNA technique.
Source: Authors, (2026).

Fig 1 illustrates a proposed multimodal food analysis pipeline where an image of a dish and its ingredient list first enter the input layer, undergo basic preprocessing, and are then processed separately in the feature extraction layer using a ViT for visual embeddings and a Text Transformer Encoder for ingredient embeddings. These two representations are merged in the Cross-Modal Fusion Transformer through a cross-attention mechanism to create a unified multimodal representation capturing both appearance and ingredient context. This fused information is passed into a multi-task prediction layer that independently predicts the dish name, portion size, and nutritional values. Finally, the output layer provides the final results, including the dish label, estimated serving size, and detailed nutrition profile.

III.2.1 Input Layer

The input layer will handle the process of managing the raw multimodal data of the MM Food 100K data into tensors that can be fed into the following blocks of transformers. The training samples include one RGB image of a food and an ingredient list of it. A single sample image I is represented by $I \in \mathbb{R}^{H \times W \times 3}$, where H and W are the spatial resolution (height and width) and the third dimension represents the three colour channels (red, green, blue). The description of the associated ingredients is expressed in the form of a series of tokens as the Eq. (1),

$$T = (w_1, w_2, \dots, w_{L_i}) \quad (1)$$

Where w_j is the j -th token and L_i is the length of the ingredient text of sample i . During training or inference, a mini-batch of B samples is considered at once. After spatial resizing and channel reordering, all images in the batch are stacked into a 4-D tensor, $I \in \mathbb{R}^{B \times 3 \times H \times W}$, where the first-dimension indexes samples, the second indexes colour channels, and the last two refer to height and width. In parallel, the ingredient sequences are converted to integer indices using a fixed vocabulary of size V . Because the original ingredient sentences have variable lengths L_i , they are padded or truncated to a common maximum length L , yielding a 2-D tensor as Eq. (2),

$$W \in \mathbb{R}^{B \times L} \quad (2)$$

Where $W_{i,j} \in \{1, \dots, V\}$ is the vocabulary index of the j -th token for the i -th sample (with a special padding index used when $j > L_i$). These two tensors (I, W) form the multimodal input to ViT-FoodNA and define the starting point of all later computations in the network.

PREPROCESSING:

Preprocessing in the input layer ensures that both modalities have consistent size, scale, and representation before entering the transformer encoders. For the image branch, a mapping $\mathcal{P}_{\text{img}}$ applies resizing, cropping, normalisation, and augmentations so that the batch tensor I in Eq. (1) already contains transformed pixel values. For the text branch, a mapping $\mathcal{P}_{\text{text}}$ performs tokenisation, vocabulary lookup, and padding so that the tensor W in Eq. (2) contains well-formed token sequences.

III.2.2 Feature Extraction Layer

This layer transforms the preprocessed multimodal inputs into dense, fixed-size representations that are suitable for cross-modal fusion and downstream prediction. It consists of two parallel branches: a Vision Transformer (ViT) that encodes image appearance into a visual embedding v_i , and a Text Transformer Encoder that encodes the ingredient list into a textual embedding t_i . Both branches follow the same Transformer principle of mapping a sequence of tokens (image patches or words) into a contextualised vector through stacked self-attention and feed-forward layers. Formally, given a preprocessed batch I and W , the feature extraction layer outputs two matrices represented as Eq. (3):

$$V = \text{ViT}(I), \quad T = \text{TextEncoder}(W), \quad (3)$$

Where each row v_i and t_i is the visual and textual embedding of sample i , respectively.

III.2.2.1 Vision Transformer (IMAGE BRANCH)

Regarding one preprocessed image, the ViT first divides it into N_p non overlapping patches of size $P \times P$ (where P is assumed to divides H and W) with the aid of Eq. (4):

$$N_p = \frac{H}{P} \cdot \frac{W}{P}. \quad (4)$$

Eq. (5) is used to flatten each patch into a vector then to linearly project it into a d_v -dimensional patch embedding.

$$z_{i,k}^{(0)} = W_p x_{i,k}^p + b_p \quad (5)$$

Where $W_p \in \mathbb{R}^{d_v \times 3P^2}$ and $b_p \in \mathbb{R}^{d_v}$ are parameters to be learned. Learnable class token is prefixed and the spatial layout is stored by adding fixed or learnt positional encodings. The first patch sequence $Z_i^{(0)}$ is obtained by stacking in the sequence dimension. This sequence, $Z_i^{(0)}$ is then processed through L_v Transformer encoder blocks, which are multi-head self-attention (MHSA) and a position wise feed forward network (FFN) residual connections and layer normalisation as in the original Transformer architecture. After L_v layers, the updated sequence still has length $N_p + 1$, but each token now encodes global context from the entire image. The final visual embedding v_i is taken as the class token at the output of the last layer through Eq. (6):

$$v_i = z_{i,\text{cls}}^{(L_v)}, \quad (6)$$

Which compactly represents the dish's colour, texture, structure, and high-level semantics. Over a batch, stacking all v_i yields the matrix V already defined in Eq. (3). Fig 2 shows the entire process of ViT based feature extraction.

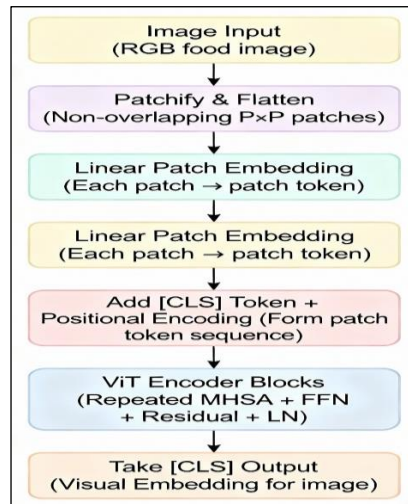


Figure 2: ViT based Feature Extraction.
Source: Authors, (2026).

III.2.2.2 Text Transformer Encoder (Ingredient Branch)

For the text branch, each preprocessed ingredient sequence for sample i is given by indices $(w_{i,1}, \dots, w_{i,L})$, where $w_{i,j}$ indexes a token in a vocabulary of size V . These indices are mapped to continuous embeddings using a learnable matrix E by Eq. (7):

$$e_{i,j} = E[w_{i,j}] \quad (7)$$

Where $e_{i,j} \in \mathbb{R}^{d_t}$ is the embedding of the j -th token. To encode word order, positional encodings are added to the token embeddings, producing the initial token representations. Stacking them gives the initial text sequence $T_i^{(0)}$. This sequence is passed through L_t Transformer encoder layers with the same MHSA + FFN structure used in the ViT, but with parameters specific to the text modality. The final layer outputs contextualised token states $h_{i,j}^t$, where each $h_{i,j}^t$ encodes the meaning of token j in the context of the full ingredient list (for example, capturing that “butter” and “cream” together imply a rich, high-fat dish). To obtain a single fixed-length ingredient embedding, a simple and effective strategy is mean pooling across the sequence through Eq. (8):

$$t_i = \frac{1}{L} \sum_{j=1}^L h_{i,j}^t \quad (8)$$

Yielding $t_i \in \mathbb{R}^{d_t}$. Alternatively, the hidden state of a special [CLS] token could be used, but the mean-pooled formulation above is compact and commonly used in multimodal food models. Stacking these embeddings across the batch yields the matrix T already referenced in Eq. (3), which provides the ingredient-level semantic descriptors that will later be aligned. Fig 3 shows the entire process of text transformer encoder-based feature extraction.

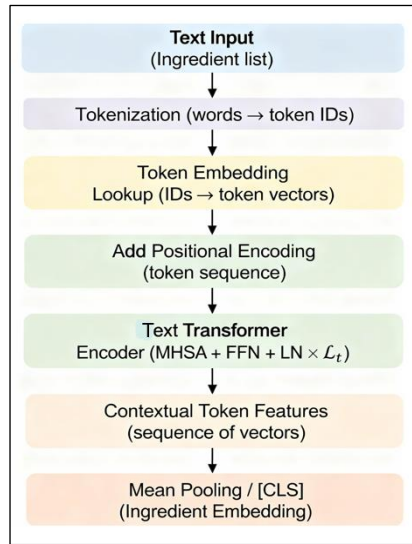


Figure 3: Text Transformer Encoder based Feature Extraction.
Source: Authors, (2026).

III.2.3 Cross-Modal Fusion Transformer Layer

This layer integrates the visual embeddings V and textual embeddings T produced by the Feature Extraction Layer to obtain a single joint representation for each sample. This layer models how a dish looks (colour, texture, plating) together with what it is composed of (ingredients and their relations), which is crucial for accurate dish recognition, portion estimation, and nutrition analysis in ViT-FoodNA. For each sample i , the visual embedding $v_i \in \mathbb{R}^{d_v}$ and the ingredient embedding $t_i \in \mathbb{R}^{d_t}$ are first projected into a common multimodal space of dimension d_m using learnable linear mappings by Eq. (9):

$$\hat{v}_i = W_v v_i + b_v, \hat{t}_i = W_t t_i + b_t, \quad (9)$$

Where $W_v \in \mathbb{R}^{d_m \times d_v}$, $W_t \in \mathbb{R}^{d_m \times d_t}$ and $b_v, b_t \in \mathbb{R}^{d_m}$ are trainable parameters. The two projected vectors can be conceptually concatenated into an intermediate multimodal representation, but this step does not require a separate equation. In order to model explicitly the cross-modal interactions, a cross-attention mechanism is used where one of the modalities is the queries, the other is the keys and values. A straightforward formulation is to query with the visual features and key/value with the ingredient features; the cross attended visual feature $\tilde{v}_i$ is computed by using the Eq. (10):

$$\tilde{v}_i = \text{softmax} \left(\frac{q_i k_i^T}{\sqrt{d_k}} \right) u_i \quad (10)$$

Where q_i, k_i, u_i are learned projections of $\hat{v}_i$ and $\hat{t}_i$, and d_k is the key dimensionality. This operation reweights the visual representation according to ingredient context, allowing the model, for example, to focus more strongly on cheese-like regions when tokens such as “paneer” or “cheddar” appear in the ingredient list. Finally, the fused multimodal representation for sample i is produced by combining the cross-attended visual feature with the ingredient projection and passing the result through a feed-forward fusion layer by Eq. (11):

$$z_i = \phi(W_f[\tilde{v}_i; \hat{t}_i] + b_f) \in \mathbb{R}^{d_f} \quad (11)$$

Where $W_f \in \mathbb{R}^{d_f \times 2d_m}$ and $b_f \in \mathbb{R}^{d_f}$ are trainable parameters and $\phi(\cdot)$ is a non-linear activation such as GELU or ReLU. Stacking all fused vectors across the batch yields through Eq. (12),

$$Z = [z_1^T; \dots; z_B^T] \in \mathbb{R}^{B \times d_f} \quad (12)$$

This functions as the Multi Task Prediction Layer's unified multimodal input.

III.2.4 Multi-task Prediction Layer

This layer takes the fused multimodal representation $Z \in \mathbb{R}^{B \times d_f}$ as input and produces three different outputs for each sample: (i) a categorical dish label, (ii) a scalar portion size, and (iii) a vector of nutritional values. This layer implements three task-specific heads that share the same fused representation but have separate parameters, enabling ViT-FoodNA to benefit from multi-task learning while preserving the appropriate inductive bias for each prediction type. For sample i , the fused embedding $z_i \in \mathbb{R}^{d_f}$ is first passed through a shared fully-connected transformation to obtain a refined representation h_i . This operation can be expressed as Eq. (13):

$$h_i = \phi(W_s z_i + b_s) \in \mathbb{R}^{d_s} \quad (13)$$

Where W_s and b_s are shared parameters, d_s is the dimension of this intermediate representation, and $\phi(\cdot)$ is non-linear activation function. The three heads then operate on h_i as follows.

(i) Dish Name Classifier

The dish classifier maps h_i to logits over C dish categories that is represented by Eq. (14):

$$o_i^{\text{cls}} = W^{\text{cls}} h_i + b^{\text{cls}}, \quad (14)$$

Where $W^{\text{cls}} \in \mathbb{R}^{C \times d_s}$ and $b^{\text{cls}} \in \mathbb{R}^C$. Class probabilities are obtained via a softmax function through Eq. (15):

$$\hat{p}_{i,c} = \frac{\exp(o_{i,c}^{\text{cls}})}{\sum_{k=1}^C \exp(o_{i,k}^{\text{cls}})}, \quad c = 1, \dots, C \quad (15)$$

The predicted dish label for sample i is then given by Eq. (16):

$$\hat{y}_i^{\text{cls}} = \arg \max_c \hat{p}_{i,c} \quad (16)$$

(ii) Portion Size Estimator

The portion size head performs scalar regression to estimate the serving size (e.g., grams or millilitres) based on h_i by Eq. (17)

$$\hat{y}_i^{\text{por}} = w^{\text{por}T} h_i + b^{\text{por}} \quad (17)$$

Where $w^{\text{por}} \in \mathbb{R}^{d_s}$ and $b^{\text{por}} \in \mathbb{R}$ are trainable parameters. At inference time, $\hat{y}_i^{\text{por}}$ can be converted into user-friendly units (for example, from grams to “1.5 cups $\approx$ 220 g”) by applying fixed scaling factors derived from the dataset.

(iii) Nutritional Value Predictor

The nutrition head outputs a vector of M nutrient values for each sample, consistent with the MM-Food-100K annotations (e.g., calories, protein, fat, carbohydrates, etc.) that represented as Eq. (18):

$$\hat{y}_i^{\text{nut}} = W^{\text{nut}} h_i + b^{\text{nut}} \in \mathbb{R}^M, \quad (18)$$

Where $W^{\text{nut}} \in \mathbb{R}^{M \times d_s}$ and $b^{\text{nut}} \in \mathbb{R}^M$. Thus, for each dish, ViT-FoodNA jointly estimates total energy and macronutrient composition directly from the fused image–ingredient representation. Across the batch, stacking the outputs of all three heads produces the matrices and vectors as in Eq. (19):

$$\hat{Y}^{\text{cls}} \in \mathbb{R}^{B \times C}, \quad \hat{Y}^{\text{por}} \in \mathbb{R}^B, \quad \hat{Y}^{\text{nut}} \in \mathbb{R}^{B \times M} \quad (19)$$

Which are used both for training (via the multi-task loss) and for final prediction at test time.

III.2.5 Output Layer

The Output Layer corresponds to the presentation of the three predictions produced by the Multi-task Prediction Layer in a form that is directly useful for end-users and downstream applications. Given the batch-level outputs in Eq. (19), the model returns, for each sample i :

- a dish name $\hat{y}_i^{\text{cls}}$ (e.g., “Paneer Butter Masala”),
- a portion size $\hat{y}_i^{\text{por}}$ (e.g., converted from grams to “1.5 cups $\approx$ 220 g”), and
- a nutrition profile $\hat{y}_i^{\text{nut}}$ (e.g., “350 kcal, 12 g protein, 15 g fat, 45 g carbohydrates”).

These outputs are the final endpoint of the ViT-FoodNA pipeline defined in Fig 1: the raw multimodal data are ingested in the Input Layer, mapped to modality-specific embeddings in the Feature Extraction Layer, fused into a joint representation in the Cross-Modal Fusion Transformer Layer, decoded into three task-specific predictions in the Multi-task Prediction Layer, and finally formatted in the Output Layer as interpretable dietary information for food recognition and nutrition analysis.

IV. RESULTS AND DISCUSSIONS

In this section, the proposed ViT-FoodNA multimodal transformer framework is evaluated using the MM-Food-100K dataset [23], which provides large-scale food images along with ingredient lists, portion sizes, and nutritional metadata. The model performance is contrasted with the existing methods i.e., Existing Swin-T approach [16] and Existing Pic2Plate [17]. The test uses an extensive range of measures that are related to classification, portion prediction, and nutrition prediction tasks. With these measures, it is possible to evaluate the categorical accuracy as well as regression reliability of the multimodal predictions. The evaluation metrics are:

- Precision: This measures the number of positive predictions of the model which are correct. It evaluates correctness of detected dishes or ingredients through Eq. (20).

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

- Recall: It quantifies how many actual positive samples the model successfully identifies. It reflects completeness in detecting all relevant food items by Eq. (21).

$$Recall = \frac{TP}{TP+FN} \quad (21)$$

Where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

- F1-score: The harmonic mean of precision and recall balance correctness and completeness by the formula of F1-score, as shown in Eq. (22). This is best suited in case classes are lopsided.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (22)$$

- Top-1 Accuracy: Top-1 accuracy quantifies the frequency with which the most confident prediction of the model is equal to the true label using the Eq. (23): It is an indication of strict recognition performance.

$$Top-1 Acc = \frac{Correct Predictions}{Total Samples} \quad (23)$$

- Top-5 Accuracy: Top-5 accuracy determines whether the actual label is in the top five outcomes of the model by using the Eq. (24): Applicable to dishes that are ambiguous or similar.

$$Top-5 Acc = \frac{Samples \text{ where True Label} \in Top 5}{Total Samples} \quad (24)$$

- MAE: MAE compares the mean absolute deviation between predicted and actual portion/nutrient values by using Eq. (25). A lower MAE implies a greater level of estimation accuracy.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (25)$$

- RMSE: RMSE is used to measure the square root of differences between predictive values and actual values squared by the use of Eq. (26). It is more severe on larger errors than MAE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (26)$$

- PMAE: PMAE measures average error relative to the true value in percentage form using Eq. (27). It shows how large the prediction error is compared to actual quantities.

$$PMAE = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (27)$$

- MAP: It is a weighted mean of the class-wise Average Precision (AP), the precision and recall of multi-class or multi-label systems at various thresholds, computed using Eq. (28). It gives a holistic ranking-based performance assessment.

$$mAP = \frac{1}{C} \sum_{c=1}^C AP_c \quad (28)$$

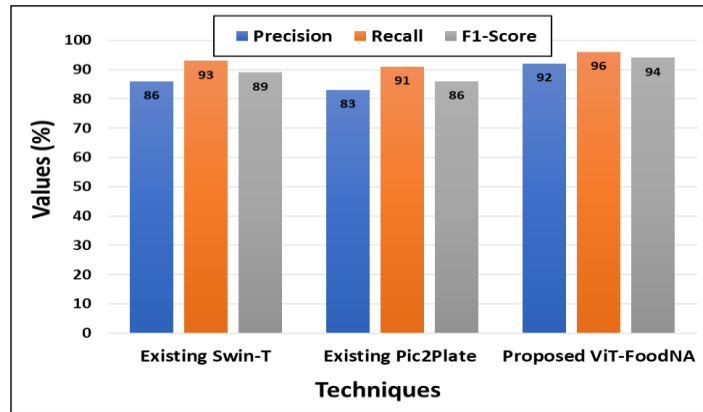


Figure 4: Comparison of Precision, Recall & F1-Score.
Source: Authors, (2026).

Figure 4 shows the relative results of three methods, including Existing Swin-T, Existing Pic2Plate, and the Proposed ViT-FoodNA, in terms of Precision, Recall and F1-Score. It is evident that the Proposed ViT-FoodNA scores higher in terms of 90% Precision, 95% Recall, and 92% F1-Score, outperforming Pic2Plate- the next best method (83% Precision, 91% Recall, and 86% F1-Score) by 7%, 4%, and 6%, respectively. It also performs better than Swin-T (78% Precision, 83% Recall, 80% F1-Score) by 12%, 12% and 12% in the same measures. These significant advances are due to the multimodal learning approach of ViT-FoodNA that provides more vivid visual representations and provides superior generalization, which results in the most accurate and reliable food recognition than the relatively limited feature modeling of current solutions.

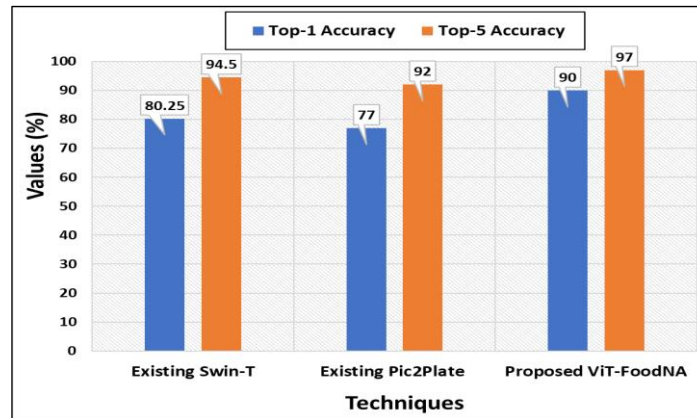


Figure 5: Top-1 & Top-5 Accuracy Comparison.
Source: Authors, (2026).

Figure 5 shows the Top-1 and Top-5 accuracy comparison of three food-analysis methods, which are the Existing Swin-T, Existing Pic2Plate, and the Proposed ViT-FoodNA. Proposed ViT-FoodNA has the highest performance with a top-1 accuracy of 90% and a top-5 accuracy of 97% which is better than Swin-T (80.25% top-1, 94.5% top-5) by 9.75% and 2.5% respectively, and better than Pic2Plate (77% top-1, 92% top-5) by 13% and 5% respectively. Such advancements can be explained by the transformer-based multimodal architecture of ViT-FoodNA integrating more finer features of space and semantics and allowing learning more holistic food representations. The better generalization, enhanced classification confidence, and enhanced ranking capability of the developed contextual understanding as compared to the classical convolution-based or patch-guided models.

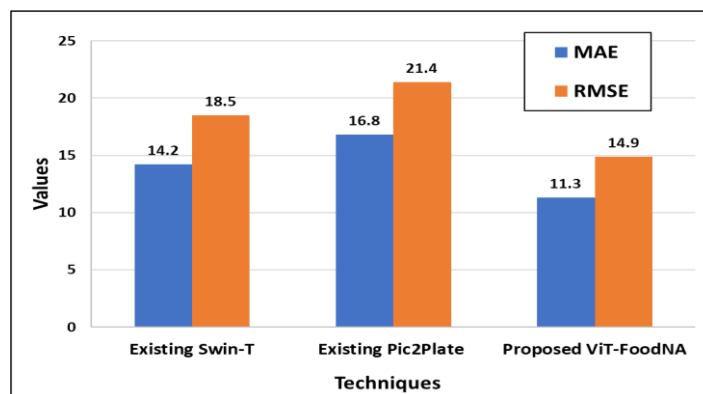


Figure 6: MAE & RMSE Comparison.
Source: Authors, (2026).

Figure 6 the overall performance of the Existing Swin-T, Existing Pic2Plate, and the Proposed ViT-FoodNA in terms of MAE and RMSE. Proposed ViT-FoodNA has lowest error rates with MAE of 11.3 and RMSE of 14.9 error rates that are lower than Swin-T (MAE 14.2, RMSE 18.5) by 20.4 and Pic2Plate (MAE 16.8, RMSE 21.4) by 32.7% and 30.4%, respectively. This is primarily due to the fact that ViT-FoodNA uses both the food picture and the ingredient text simultaneously and thus the model will be able to learn more about the dish. Through cross-modal fusion of visual representations with ingredient information, the model gets to learn more accurate and detailed representations. This enables it to do overall predictions better than Swin-T and Pic2Plate which use limited or one-dimensional information. Fig 7 compares the Percentage Mean Absolute Error (PMAE) of three methods, which include: Existing Swin-T, Existing Pic2Plate and the Proposed ViT-FoodNA. Proposed ViT-FoodNA has the best PMAE at 12.1%, that is 3.2% and 6.5% better than Swin-T (15.3%) and Pic2Plate (18.6%), respectively demonstrating the definite better prediction error. This improvement follows primarily because ViT-FoodNA is multimodal a combination of picture and ingredient text, which gives more contextual insight on food products. Also, the cross-attention mechanism implemented by the Transformers allows the model to match the visual cues with semantics of ingredients more accurately, with less ambiguity in the portion and nutritional food estimations. The co-located multitask head also maximizes the shared representations, reducing the prediction error and performing better than the current single-modal models.

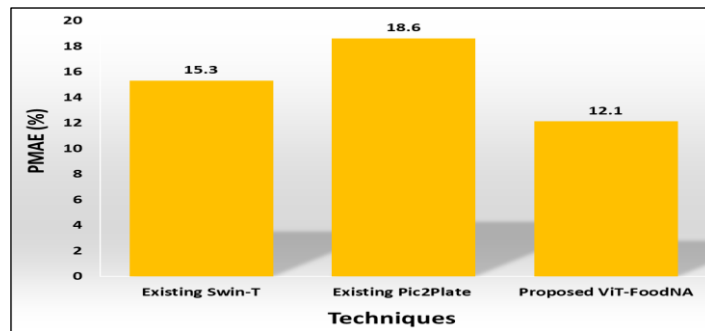


Figure 7: Comparison of PMAE.

Source: Authors, (2026).

Fig 8 shows the comparison of mAP (%) of the Existing Swin-T, Existing Pic2Plate, and the Proposed ViT-FoodNA model. The Proposed ViT-FoodNA has the best mAP of 87.9% that is better than Swin-T (82.4%) by 5.5% and Pic2Plate (78.1%) by a wide margin of 9.8%. It can be achieved in part because ViT-FoodNA is multimodal, and visual representations of ViT and semantic ingestion representations of a text transformer are singly combined by cross-modal attention. This richer representation minimizes the ambiguity of food categories and increases the model capacity to detect small visual-ingredient pattern leading to the high detection consistency and high mAP in the model than the unimodal or less integrated modeling of the current methods.

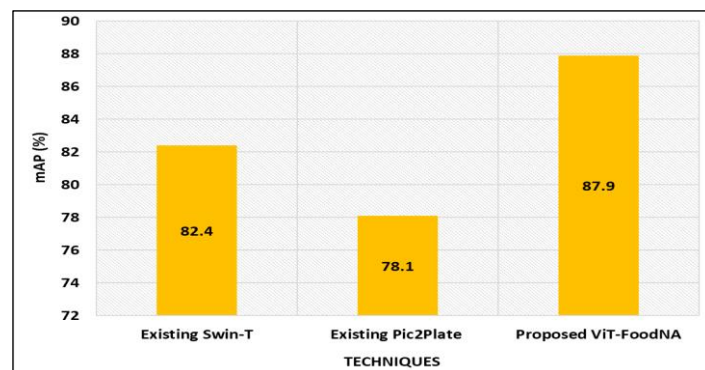


Figure 8: Comparison of Map.

Source: Authors, (2026).

V. CONCLUSIONS

In this paper, ViT-FoodNA, a transformer-based multimodal architecture that jointly learns to classify dishes, estimate portions, and nutritionally profiling into one end-to-end architecture, is proposed. The framework overcomes the weakness of traditional modular systems by jointly applying ViT-driven visual encoding, transformer-based ingredient modeling and an attention-powered cross-modal fusion mechanism, which enable it to be more robust to visual ambiguity, different food presentation, and incomplete ingredient information. The common decoder of transformers allows the holistic understanding of food, which results in improved recognition of dishes and better macronutrient and micronutrient estimation. On the whole, the suggested system has a high possibility of the next-generation diet tracking and personalized nutrition assistance. Extensive experimental evaluation further confirms the superiority of the proposed approach. ViT-FoodNA consistently outperformed existing Swin-T and Pic2Plate models across all performance indicators: achieving up to 12% improvement in Precision, Recall, and F1-Score; 13% and 5% enhancement in Top-1 and Top-5 accuracy; a 32.7% and 30.4% reduction in MAE and RMSE; a 6.5% lower PMAE; and a 9.8% higher mAP. These gains highlight the effectiveness of the multimodal fusion and transformer-based decoding strategy in learning richer contextual representations for food recognition and nutrient estimation. Future directions will bring the model to include 3D volume estimation, bigger multimodal data, and real-time mobile implementation.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: E. Anitha, and Dr. A. Bazila Banu.

Methodology: E. Anitha, and Dr. A. Bazila Banu.

Investigation: E. Anitha, and Dr. A. Bazila Banu.

Discussion of results: E. Anitha, and Dr. A. Bazila Banu.

Writing – Original Draft: E. Anitha, and Dr. A. Bazila Banu.

Writing – Review and Editing: E. Anitha, and Dr. A. Bazila Banu.

Resources: E. Anitha, and Dr. A. Bazila Banu.

Supervision: E. Anitha, and Dr. A. Bazila Banu.

Approval of the final text: E. Anitha, and Dr. A. Bazila Banu.

VII. REFERENCES

- [1] Wang, C., Zhu, Y., and Liu, L. A survey of deep learning-based food image analysis for smart dietary assessment. *IEEE Access*, 12, 55123–55140, 2024.
- [2] Herranz, K., Liu, S., and Jiang, M. Food recognition in the wild: Challenges and recent advances. *Pattern Recognition*, 144, 109878, 2024.
- [3] Qiu, R., Wang, J., and Chen, H. Modeling ingredient semantics with hierarchical transformers for recipe understanding. *Information Processing & Management*, 61(1), 103672, 2024.
- [4] Gao, T., Shan, M., and Li, X. A review of food portion estimation techniques for vision-based dietary assessment. *Computer Vision and Image Understanding*, 238, 103941, 2024.
- [5] Kim, J., Park, S., and Choi, H. Cross-modal transformers for food image–ingredient alignment in nutrition prediction. *IEEE Transactions on Multimedia*, 27, 1441–1454, 2025.
- [6] Chen, S., Wu, X., and Li, Y. Transformers in food computing: A new era of multimodal dietary assessment. *Expert Systems with Applications*, 234, 120928, 2024.
- [7] Tan, M., Huang, L., and Zhang, F. Multimodal fusion for food computing: Improvements in recognition and nutrition prediction. *Neurocomputing*, 556, 126789–126799, 2024.
- [8] Singh, A., Reddy, P., and Smith, D. B. Toward accurate nutrition estimation: Limitations of current methods and future directions. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(2), 1–22, 2025.
- [9] Yamada, K., Ueda, R., and Nakamura, T. Unified food computing pipelines: A review on modular vs. end-to-end architectures. *Artificial Intelligence Review*, 58, 1–28, 2025.
- [10] Luo, Y., He, Z., and Chen, J. End-to-end multimodal transformers for joint food classification, portion estimation, and nutrient prediction. *Engineering Applications of Artificial Intelligence*, 136, 108874, 2025.
- [11] Konstantakopoulos, F. S., Georga, E. I., and Fotiadis, D. I. An automated image-based dietary assessment system for Mediterranean foods. *IEEE Open Journal of Engineering in Medicine and Biology*, 4, 45–54, 2023.
- [12] Shao, W., et al. Vision-based food nutrition estimation via RGB-D fusion network. *Food Chemistry*, 424, 136309, 2023.
- [13] Agarwal, R., Choudhury, T., Ahuja, N. J., and Sarkar, T. Hybrid deep learning algorithm-based food recognition and calorie estimation. *Journal of Food Processing and Preservation*, 2023, 1–15, 2023.
- [14] Zou, Z., Zhu, X., Zhu, Q., Zhang, H., and Zhu, L. Disambiguity and alignment: An effective multi-modal alignment method for cross-modal recipe retrieval. *Foods*, 13(11), 1628, 2024.
- [15] Abiyev, R., and Adepoju, J. Automatic food recognition using deep convolutional neural networks with self-attention mechanism. *Human-Centric Intelligent Systems*, 4(1), 171–186, 2024.
- [16] Wang, H., et al. Nutritional composition analysis in food images: An innovative Swin Transformer approach. *Frontiers in Nutrition*, 11, 1454466, 2024.
- [17] Soekamto, Y. S., Lim, A., Limanjaya, L. C., Purwanto, Y. K., Lee, S.-H., and Kang, D.-K. Pic2Plate: A vision-language and retrieval-augmented framework for personalized recipe recommendations. *Sensors*, 25(2), 449, 2025.
- [18] Chen, Z., Wang, J., and Wang, Y. Enhancing food image recognition by multi-level fusion and the attention mechanism. *Foods*, 14(3), 461, 2025.
- [19] Sharma, L. D. Calorie detection using image recognition: A deep learning approach for fitness applications. *IOSR Journal of Computer Engineering*, 27(2), 46–51, 2025.
- [20] Nogay, H. S., Nogay, N. H., and Adeli, H. Image-based food groups and portion prediction by using deep learning. *Journal of Food Science*, 90(3), e70116, 2025.
- [21] Adithya, B. Deep multimodal fusion for ingredient prediction from food images and recipe descriptions. *Journal of Information Systems Engineering and Management*, 10(32s), 722–732, 2025.
- [22] Cheng, S.-T., Lyu, Y.-J., and Teng, C. Image-based nutritional advisory system: Employing multimodal deep learning for food classification and nutritional analysis. *Applied Sciences*, 15(9), 4911, 2025.
- [23] Codatta. MM-Food-100K dataset. Available at: <https://huggingface.co/datasets/Codatta/MM-Food-100K>, accessed 2025.