frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 14 October 2024
DOI 10.3389/fnut.2024.1454466

Check for updates

OPEN ACCESS

EDITED BY
Tarun Belwal,
Texas A&M University, United States

REVIEWED BY
Maria Valero,
Kennesaw State University, United States
Xizhe Fu,
Shihezi University, China
Monika Thakur,
Amity University, India

*CORRESPONDENCE
Hui Wang
✉ wanghui20240417@163.com

RECEIVED 25 June 2024
ACCEPTED 12 August 2024
PUBLISHED 14 October 2024

CITATION
Wang H, Tian H, Ju R, Ma L, Yang L, Chen J and Liu F (2024) Nutritional composition analysis in food images: an innovative Swin Transformer approach.
Front. Nutr. 11:1454466.
doi: 10.3389/fnut.2024.1454466

Nutritional composition analysis in food images: an innovative Swin Transformer approach

Hui Wang[1]*, Haixia Tian[2], Ronghui Ju[1], Liyan Ma[3], Ling Yang[1], Jingyao Chen[1] and Feng Liu[4]

[1]College of Food and Biological Engineering, Beijing Vocational College of Agriculture, Beijing, China, [2]China Tea Technology Co., Ltd., Beijing, China, [3]College of Food Science and Nutrition Engineering, China Agricultural University, Beijing, China, [4]Beijing Sanyuan Foods Co., Ltd., Beijing, China

Accurate recognition of nutritional components in food is crucial for dietary management and health monitoring. Current methods often rely on traditional chemical analysis techniques, which are time-consuming, require destructive sampling, and are not suitable for large-scale or real-time applications. Therefore, there is a pressing need for efficient, non-destructive, and accurate methods to identify and quantify nutrients in food. In this study, we propose a novel deep learning model that integrates EfficientNet, Swin Transformer, and Feature Pyramid Network (FPN) to enhance the accuracy and efficiency of food nutrient recognition. Our model combines the strengths of EfficientNet for feature extraction, Swin Transformer for capturing long-range dependencies, and FPN for multi-scale feature fusion. Experimental results demonstrate that our model significantly outperforms existing methods. On the Nutrition5k dataset, it achieves a Top-1 accuracy of 79.50% and a Mean Absolute Percentage Error (MAPE) for calorie prediction of 14.72%. On the ChinaMartFood109 dataset, the model achieves a Top-1 accuracy of 80.25% and a calorie MAPE of 15.21%. These results highlight the model's robustness and adaptability across diverse food images, providing a reliable and efficient tool for rapid, non-destructive nutrient detection. This advancement supports better dietary management and enhances the understanding of food nutrition, potentially leading to more effective health monitoring applications.

KEYWORDS

nutrient recognition, EfficientNet, Swin Transformer, Feature Pyramid Network, deep learning, non-destructive detection

1 Introduction

The increasing concern for healthy diets and food quality has made the detection and analysis of food nutrients a critical research direction (1). Nutrients such as proteins, fats, carbohydrates, vitamins, and minerals are essential components for maintaining human health, and their intake and proportions have a direct impact on overall well-being (2). Therefore, accurate detection of nutrients in food is vital for formulating scientific dietary plans and ensuring food safety (3). However, traditional methods for nutrient detection often rely on chemical analysis and destructive sampling, which are time-consuming and complex, limiting their widespread application (4). In recent years, to achieve more efficient and convenient nutrient detection, computer vision and deep learning technologies have been gradually introduced to this field, offering a non-destructive solution. Currently, deep learning has made significant progress in the field of computer vision, especially in image classification, object detection, and feature extraction, demonstrating powerful capabilities (5). These technologies provide new solutions for the detection and analysis of food nutrients. Researchers can utilize deep learning models to extract and classify features from food images, enabling rapid, non-destructive nutrient detection. Deep learning applications in food nutrient detection benefit significantly from its capabilities in automatic feature extraction, precise classification, end-to-end learning, and data augmentation (6). Deep learning models can automatically learn effective features from large datasets, reducing the need for manual intervention and improving the accuracy and efficiency of feature extraction. Additionally, these models excel in image classification tasks, capable of handling complex image data and capturing subtle feature differences, thus achieving high classification accuracy. By training in an end-to-end manner, from raw image input to nutrient output, the process is greatly simplified with no manual intervention required. Furthermore, through data augmentation and transfer learning, deep learning models can be trained on limited datasets and extended to more food categories, enhancing the models' generalization capabilities (7). Despite the immense potential of deep learning in food nutrient detection, challenges remain. Issues such as model robustness and interpretability, diversity, and quality of datasets require further research and optimization (8). The primary aim of this study is to develop an efficient and accurate method for detecting food nutrients, providing a non-destructive, rapid solution for food quality assessment and dietary monitoring. By introducing advanced deep learning technologies, we hope to overcome the current challenges in nutrient detection and advance the development of this field.

In recent years, with the rapid development of deep learning technology, numerous researchers have focused on applying it to the identification and analysis of food nutrients, achieving remarkable results. In one related study, researchers used Convolutional Neural Networks (CNNs) to identify and classify food images to infer their nutrient content (9). The background of this study highlighted the high cost and time-consuming nature of traditional methods that rely on chemical analysis, whereas image recognition technology can provide a fast and non-destructive solution. This study utilized a pre-trained ResNet model to extract features from food images and employed fully connected layers for classification, achieving high classification accuracy. However, this method exhibited certain limitations due to its heavy reliance on data, particularly under varying lighting conditions and different shooting angles. Another study employed deep learning models for food image segmentation and feature extraction to more accurately identify the proportions of different food components (10). The researchers combined U-Net for image segmentation and VGG16 for feature extraction, accurately separating different component regions through the segmentation network and performing further feature analysis on these regions. This method excelled in improving the fine-grained analysis of nutrient detection, significantly enhancing the model's detection accuracy. However, the segmentation accuracy decreased when dealing with multi-component mixed foods, affecting the overall detection performance. In the third study, researchers proposed a multi-modal deep learning method for food nutrient recognition. The background of this study noted that single image information might not sufficiently describe the nutrient content of food, hence the incorporation of image, text descriptions, and nutritional label information (11). The researchers adopted a multi-modal model that combines CNNs and Long Short-Term Memory (LSTM) networks, extracting visual features through the image network, processing food description information through the text network, and performing feature fusion using an attention mechanism. Experimental results demonstrated high accuracy and robustness in multi-modal information fusion. However, the complexity of acquiring and processing multi-modal data, along with a cumbersome data preparation process, limited its application scope. The fourth related work utilized Generative Adversarial Networks (GANs) to generate high-quality food image data, enhancing the diversity and robustness of model training. This study addressed the issue of insufficient food image data, which limits the performance improvement of deep learning models (12). Researchers generated realistic food images through GANs and combined them with actual data for training, improving the model's generalization ability and detection accuracy. This approach yielded excellent detection results on multiple public food datasets. However, the generated images differed from real images in certain details, affecting the accuracy of some feature extractions. Despite the significant progress made in food nutrient detection through these studies, several common issues persist. Firstly, the quality and diversity of data remain critical factors limiting model performance, particularly in the face of complex food components where robustness and accuracy need improvement. Secondly, the complexity of multi-modal information fusion and processing increases the difficulty of data preparation and model training. Additionally, the lack of model interpretability and transparency poses challenges in gaining user trust and acceptance in practical applications. In summary, existing research has achieved certain successes in the rapid and non-destructive detection of food nutrients, but there is still room for improvement in data processing, model robustness, and interpretability. This study aims to overcome the limitations of current methods by introducing more efficient network structures and innovative feature fusion methods, further advancing the technology for food nutrient identification and analysis.

Based on the shortcomings of the aforementioned work, we propose a new model that combines EfficientNet and Swin Transformer. The aim of this new model is to address the existing methods' deficiencies in data dependency, segmentation accuracy, multi-modal data processing complexity, and the detail accuracy of generated images. Our model consists of three main components: the EfficientNet backbone network, the Swin Transformer module, and the Feature Pyramid Network (FPN) fusion module. The EfficientNet backbone network is used for efficiently extracting low-level features from food images, characterized by high parameter utilization and robust feature extraction capabilities. The Swin Transformer module captures long-range dependencies within the images, further enhancing the quality of feature representation. The FPN performs deep fusion of the extracted features, enhancing the model's classification performance by conducting attention calculations across different feature subspaces. This model has significant advantages in addressing existing issues. Firstly, by combining EfficientNet and Swin Transformer, the model's robustness to different lighting conditions and shooting angles is improved. Secondly, FPN enhances the effectiveness of feature fusion, improving segmentation and classification accuracy. Lastly, the innovative network structure and feature fusion methods simplify the data processing workflow, enhancing the model's generalization ability and detection accuracy. Additionally, our model demonstrates improvements in data processing by efficiently handling high-dimensional data and providing robust feature extraction. The integration of Swin Transformer improves model robustness by effectively capturing long-range dependencies, while FPN enhances interpretability through multi-scale feature fusion, making the model's decision-making process more transparent.

- We propose a new model that combines EfficientNet and Swin Transformer. This model integrates EfficientNet's efficient feature extraction capabilities and Swin Transformer's ability to capture long-range dependencies, using FPN for deep feature fusion, significantly improving the accuracy of food nutrient recognition and classification.
- We have developed a rapid, non-destructive method for detecting food nutrients. By leveraging advanced computer vision and deep learning technologies, this method enables quick analysis of food images and precise detection of nutrients, significantly reducing detection time compared to traditional chemical analysis methods and avoiding sample destruction.
- Our model has made significant progress in enhancing robustness and generalization capabilities. By introducing the FPN and innovative feature fusion strategies, the model demonstrates stronger robustness when handling various lighting conditions, shooting angles, and complex food components. It also shows excellent generalization across multiple food datasets, validating its effectiveness in practical applications.

2 Related work

2.1 Deep learning approaches for food image recognition and ingredient segmentation

In the field of food image recognition and ingredient segmentation, deep learning methods have demonstrated remarkable potential and broad application prospects. Traditional image recognition methods rely on manual feature extraction, which often struggles to handle complex food images, especially when dealing with diverse ingredients with varying shapes and colors (13). With the development of deep learning technologies, models such as Convolutional Neural Networks (CNNs) have shown significant advantages in processing high-dimensional data, bringing revolutionary advancements to food image recognition and ingredient segmentation (14, 15).

Deep learning-based methods for food image recognition have emerged prominently. As one of the core models in deep learning, CNNs utilize hierarchical structures to progressively extract low-level to high-level features of images, achieving remarkable results in image classification tasks (16). For instance, models such as Deep Residual Networks (ResNet) and Dense Convolutional Networks (DenseNet) have demonstrated outstanding performance on large-scale image classification datasets like ImageNet (17, 18). These models are capable of not only effectively recognizing single food categories but also handling complex scenes with mixed food items, achieving simultaneous recognition of multiple ingredients through multi-label classification techniques (19). In ingredient segmentation, deep learning shows strong capabilities. Semantic segmentation networks (e.g., U-Net and SegNet) and instance segmentation networks (e.g., Mask R-CNN) achieve precise segmentation of different ingredients in images through pixel-level classification (20, 21). These models can distinguish between food and background and further refine segmentation down to different types of ingredients. Particularly, Mask R-CNN, which combines object detection and instance segmentation, can accurately locate and segment each ingredient in complex backgrounds, providing reliable data for subsequent nutritional analysis and recipe recommendations (22).

Deep learning models based on attention mechanisms have also been applied in food image recognition and ingredient segmentation. Attention mechanisms, by assigning different weights to different regions of an image, can highlight important features and improve the recognition and segmentation accuracy of models (23). For example, models based on Transformer architectures, such as Vision Transformer (ViT) and Swin Transformer, achieve global feature extraction and interaction through self-attention mechanisms, overcoming the limitations of traditional CNNs in capturing long-range dependencies and global information (24). Shao et al. (25) introduced the Swin-Nutrition model, leveraging the Swin Transformer for nutrient analysis. This model demonstrated significant improvements in accuracy compared to traditional methods but faced challenges related to data dependency and generalization. Our proposed model integrates EfficientNet for feature extraction, providing a different approach to capturing image features. While Swin-Nutrition excels in global feature extraction through the Swin Transformer, our model focuses on efficient feature extraction and multi-scale feature fusion using EfficientNet and the FPN. Furthermore, some studies have proposed multi-task learning (MTL) methods to meet the specific needs of food image recognition and ingredient segmentation tasks (26). By jointly learning multiple related tasks, these methods enhance the generalization ability and recognition accuracy of models. For instance, modeling food classification, ingredient segmentation, and calorie estimation simultaneously not only improves the performance of each task but also reduces the computational cost of the model.

In summary, the application of deep learning methods in food image recognition and ingredient segmentation has greatly promoted the development of intelligent food analysis systems. In the future, with the further advancement of deep learning technologies and the integration of emerging artificial intelligence technologies such as Generative Adversarial Networks (GANs) and Graph Neural Networks (GNNs), the accuracy and efficiency of food image recognition and ingredient segmentation will be further enhanced, bringing more innovative applications to fields such as food safety, nutritional analysis, and personalized diet recommendations (27, 28).

2.2 Advanced applications of computer vision in the food industry

The application of computer vision technology in the food industry is rapidly expanding, becoming a significant driving force in the field. By leveraging deep learning and other advanced computer vision technologies, the food industry has achieved substantial improvements in efficiency and accuracy across various aspects such as production, quality control, and supply chain management.

Firstly, automation in food production is one of the most prominent applications of computer vision in the food industry. Computer vision systems are widely used on food processing lines for automatic detection and classification of food products (29). These systems capture food images using high-speed cameras and perform real-time analysis using deep learning algorithms, accurately identifying the type, shape, and color of food items to enable automated sorting and packaging. For instance, the application of Convolutional Neural Networks (CNNs) on food production lines has significantly improved production efficiency and product consistency (30). In quality control, computer vision technology also plays a crucial role. Traditional manual inspection is not only inefficient but also prone to human error. Computer vision systems can achieve real-time monitoring and automatic detection of food quality. Researchers have developed image recognition systems based on deep learning that can identify defects, discoloration, and contaminants on food surfaces, ensuring high consistency and safety of food quality (31). For example, using Generative Adversarial Networks (GANs) to generate high-quality training data can further enhance model robustness and detection accuracy. Moreover, the application of computer vision technology in food supply chain management is continually expanding. Automatic identification and tracking of food images enable comprehensive monitoring of the food supply chain, thereby increasing transparency and efficiency. By integrating with Internet of Things (IoT) technology, food companies can obtain and analyze data from production to sales in real-time, optimizing inventory management and distribution processes. For instance, combining image recognition with blockchain technology ensures traceability and anti-counterfeiting of food products, enhancing consumer confidence in food safety (32). Another notable trend is the rise of personalized diets and nutrition management. Computer vision technology can help consumers scan food items using smartphone cameras to obtain detailed nutritional information and health advice (33). This application not only assists individuals in better managing their diets but also supports medical institutions in formulating personalized nutrition plans for patients. For example, combining deep learning with big data analysis, personalized food recommendation systems can suggest suitable foods and recipes based on users' dietary preferences and health conditions.

Despite the significant progress made by computer vision technology in the food industry, several challenges remain. For instance, acquiring and annotating high-quality food image datasets is costly, and the robustness of models under different environments and conditions needs further improvement. Additionally, the diversity and complexity of food images pose challenges to algorithm design and optimization (34). Future research will continue to address these issues, promoting the broader application and deeper development of computer vision in the food industry through technological innovation and interdisciplinary collaboration. In summary, the application of computer vision technology in the food industry has already shown immense potential, achieving significant results in production automation, quality control, supply chain management, and personalized diets. With continuous technological advancements and deeper applications, computer vision is poised to play an increasingly important role in the food industry, driving the industry's intelligent and digital transformation.

2.3 Overview of research and applications in food nutritional assessment

Food nutritional assessment has evolved significantly with the advancement of technology and methodologies. Early approaches primarily relied on manual documentation and chemical analysis, which, while accurate, were time-consuming and labor-intensive (35). The integration of machine learning and artificial intelligence has brought transformative changes to this field. Machine learning algorithms, such as regression models and clustering algorithms, analyze extensive datasets of food items and their nutrient profiles to predict nutritional content with high accuracy (36). This method has proven essential for rapid and accurate dietary planning and health management.

Moreover, the development of portable, non-invasive devices for real-time nutrient analysis has marked a significant leap forward. Devices employing spectroscopy and sensor technology can measure macronutrients and some micronutrients directly from food samples, providing immediate feedback (37). These advancements are particularly beneficial in clinical settings and for individuals managing chronic conditions like diabetes, where timely and precise nutritional information is crucial. Blockchain technology has also emerged as a critical innovation, ensuring the accuracy and transparency of nutritional information. By securely recording and verifying the provenance and nutritional content of food products, blockchain enhances trust and accountability within the food supply chain (38). This is especially useful for validating claims related to organic or fortified foods, ensuring consumers have access to reliable nutritional data. Advancements in bioinformatics and computational biology have expanded the understanding of nutritional genomics. Researchers can now identify gene-diet interactions and their impact on health by analyzing genetic data, leading to the development of nutrigenetic profiles (39). These profiles provide personalized dietary recommendations aimed at preventing or managing specific health conditions, offering valuable insights into the role of diet in metabolic disorders and other chronic illnesses. The integration of big data analytics with nutritional epidemiology has provided deeper insights into population-level dietary patterns and their health implications (40). Large-scale studies using data from national health surveys, electronic health records, and wearable devices allow researchers to identify trends and associations between diet and health outcomes. This informs public health strategies and dietary guidelines, contributing to interventions aimed at improving nutritional status and reducing diet-related diseases. In summary, the convergence of machine learning, portable sensing technologies, blockchain, bioinformatics, and big data analytics has significantly advanced food nutritional assessment. These innovations have led to more precise, efficient, and personalized approaches, transforming how nutritional health is monitored and managed, both at individual and population levels.

3 Method

3.1 Overview of our network

To address the shortcomings of existing food nutrient recognition methods, we propose a new model based on deep learning. This model combines the strengths of EfficientNet and Swin Transformer, utilizing a Feature Pyramid Network (FPN) for deep feature fusion to enhance classification performance and accuracy. EfficientNet serves as the feature extractor, responsible for extracting low-level features from the input food images. With its efficient parameter utilization and excellent feature extraction capabilities, EfficientNet ensures high accuracy with fewer parameters. The Swin Transformer module captures long-range dependencies within the images. Through its sliding window and hierarchical structure, the Swin Transformer effectively handles high-resolution images, improving the quality of feature representation. The Feature Pyramid Network (FPN) performs deep fusion of the extracted features, calculating attention across different feature subspaces to capture more information and further enhance classification performance. In constructing the network, we first standardize the public food image dataset, adjusting image size and normalization to meet the model's input requirements. Next, we use the EfficientNet model as the backbone network to extract features from the input images, obtaining low-level information. These features are then input into the Swin Transformer module, which extracts high-level features layer by layer, capturing long-range dependencies within the images. The features extracted by the Swin Transformer are subsequently input into the FPN, which fuses information from various feature subspaces to obtain more expressive feature representations. Finally, the fused features are classified through fully connected layers, enabling the model to classify food types and nutrients based on feature representations.

During model training, a multi-task loss function is used to balance the losses of five sub-tasks, including calories, mass, fat, carbohydrates, and protein. The proposed model demonstrates several advantages and innovations. Firstly, the efficient feature extraction capability of EfficientNet ensures high accuracy with fewer parameters. Secondly, the Swin Transformer, through its sliding window and hierarchical structure, effectively captures long-range dependencies within images, enhancing the quality of feature representation. The FPN, through attention calculations in different feature subspaces, fuses more information and improves classification performance. Additionally, the model can handle various types of food images, demonstrating strong adaptability and generalization ability. In summary, this study significantly enhances the accuracy and efficiency of food nutrient recognition by innovatively combining EfficientNet and Swin Transformer and introducing the FPN. The model not only achieves rapid, non-destructive detection of food components but also provides reliable technical support for food quality assessment and healthy diet monitoring, offering broad application potential and societal value. The overall architecture of the proposed model is shown in Figure 1.

3.2 EfficientNet feature extraction

EfficientNet is a new convolutional neural network architecture proposed by the Google research team, aimed at achieving efficient parameter utilization and excellent feature extraction capabilities by comprehensively considering network depth, width, and resolution. The core idea is to use Compound Scaling to simultaneously scale the network's depth, width, and input image resolution to maximize model performance given fixed computational resources (41). Specifically, EfficientNet finds an optimal balance by adjusting the number of convolutional layers, the number of channels per layer, and the input image size, allowing the model to achieve higher accuracy with fewer parameters. Figure 2 shows the structure and working principles of the EfficientNet model. The following are some key mathematical formulas in EfficientNet and their explanations.

The compound scaling method can be described by the following equation:

$$D = \alpha^d, \quad W = \beta^d, \quad R = \gamma^d \tag{1}$$

where α, β, and γ are constants; d is the compound coefficient.

The constraint for the compound scaling method is given by:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{2}$$

where α, β, and γ are constants ensuring balanced scaling; The selection of the value 2 helps maintain a balanced scaling of the network's depth, width, and resolution. This balance ensures that the model scales efficiently across different dimensions without disproportionately increasing the computational burden. The value was empirically validated to provide an optimal trade-off between model performance and computational efficiency.

In EfficientNet, the Mobile Inverted Bottleneck Convolution (MBConv) module first applies depthwise separable convolution:

$$Y = \mathrm{DepthwiseConv}(X, W_d) \tag{3}$$

where X is the input feature map; W_d is the weight of the depthwise convolution; Y is the output feature map.

This is followed by pointwise convolution:

$$Z = \mathrm{PointwiseConv}(Y, W_p) \tag{4}$$

where Y is the output of the depthwise convolution; W_p is the weight of the pointwise convolution; Z is the resulting feature map.

In the Squeeze-and-Excitation (SE) block, the channel-wise recalibration of feature maps is computed as follows:

$$S = \sigma(\mathrm{FC2}(\mathrm{ReLU}(\mathrm{FC1}(Z)))) \tag{5}$$

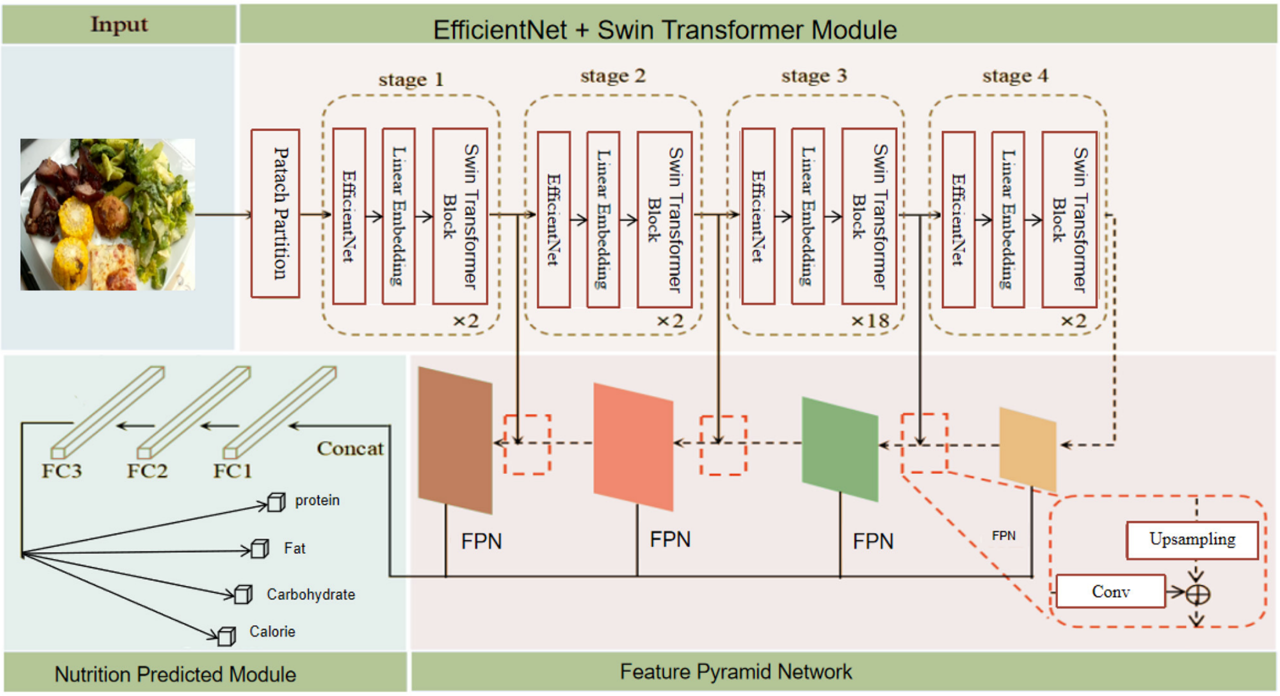


FIGURE 1
Overall structure of the proposed model.The model includes the following components: Patch Partition, EfficientNet, Swin Transformer Block, Feature Pyramid Network, Upsampling and Convolution layers, and Fully Connected layers (FC1, FC2, FC3). The model outputs predictions for protein, fat, carbohydrate, and calorie content.

where Z is the input feature map; FC1 and FC2 are fully connected layers; ReLU is the Rectified Linear Unit activation function; σ is the sigmoid function; S is the scale vector.

The recalibrated feature map is then scaled:

$$Z' = S \cdot Z \tag{6}$$

where S is the scale vector; Z is the input feature map; Z' is the scaled feature map.

Through these formulas, EfficientNet effectively extracts and processes features from the input images, providing high-quality feature representations for subsequent network modules.

EfficientNet is used for feature extraction due to its efficient parameter utilization and superior feature extraction capabilities. This helps in reducing the complexity of data processing and enhances the quality of extracted features, ensuring more reliable inputs for subsequent stages. In our proposed food nutrient recognition model, we integrate the corresponding blocks of EfficientNet into each stage. This approach allows us to leverage EfficientNet's advanced feature extraction capabilities while maintaining computational efficiency. Each stage utilizes specific blocks from EfficientNet that are most suitable for the task at hand, ensuring that our model benefits from EfficientNet's strengths without the overhead of processing the entire network. These high-quality features provide the necessary input for the subsequent Swin Transformer module and the FPN, enabling the entire model to more accurately identify food types and nutrients. Compared to other traditional convolutional neural networks (such as ResNet and VGG), EfficientNet significantly reduces computational resource consumption while maintaining high feature extraction capabilities, allowing our model to improve training and inference efficiency while ensuring high accuracy. With the high-quality features extracted by EfficientNet, our model can more accurately identify different food components and calculate their corresponding nutritional values. This is crucial for achieving non-destructive nutrient detection, enhancing the efficiency and accuracy of food detection. Additionally, the application of EfficientNet effectively improves the generalization ability of our model across different datasets, enhancing the model's robustness when handling various types of food images. This is of great significance for practical applications in food classification and nutrient recognition.

3.3 Swin Transformer module

Swin Transformer is a novel vision transformer designed to efficiently handle high-resolution images through a hierarchical structure and sliding window mechanism. Unlike traditional convolutional neural networks (CNNs), Swin Transformer captures long-range dependencies in images via self-attention mechanisms, thereby enhancing the model's feature representation capabilities while maintaining computational efficiency (42). The basic unit of Swin Transformer is the Window-based Multi-Head Self-Attention (W-MSA). In W-MSA, images are divided into non-overlapping windows, and self-attention is computed independently within each window. This approach reduces

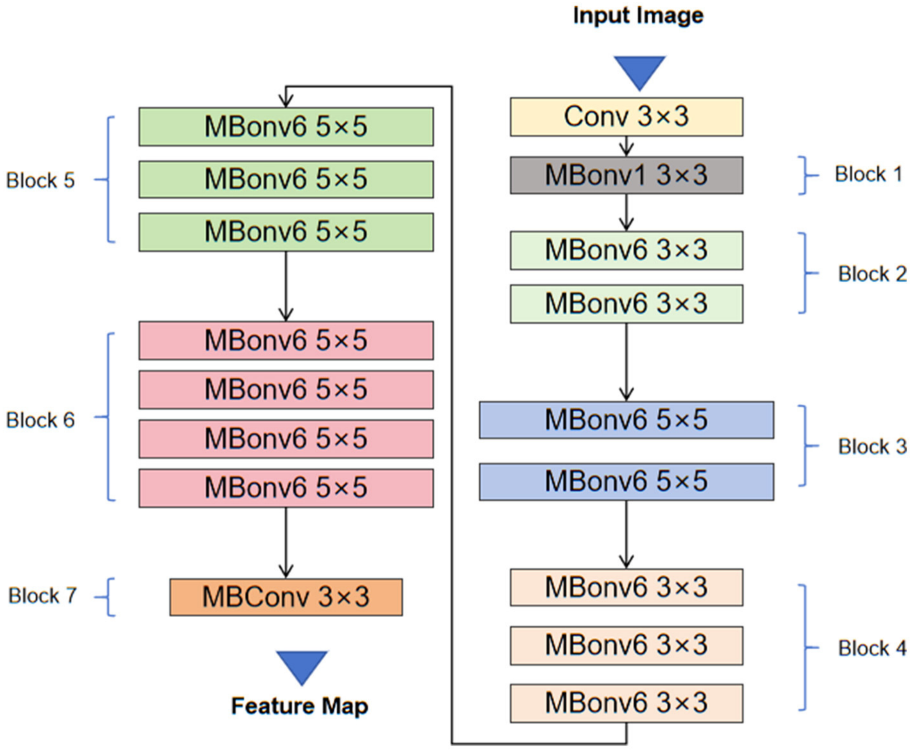


FIGURE 2
The architecture of EfficientNet.

computational complexity while preserving local features. Then, the Shifted Window Mechanism alternates window positions between different layers, enabling the model to capture long-range dependencies across windows. Figure 3 illustrates the structure and working principles of the Swin Transformer model.

In our model, the Swin Transformer is used to further process the low-level features extracted by EfficientNet, capturing higher-level features and global information to enhance the model's feature representation capabilities and classification performance. The hierarchical structure and sliding window mechanism of the Swin Transformer allow the model to extract features at different scales, improving the ability to capture both image details and global information. Compared to the global self-attention mechanism, the window-based multi-head self-attention significantly reduces computational complexity, making the model more efficient when handling high-resolution images. The integration of the Swin Transformer in our model enhances robustness by capturing long-range dependencies within food images. This allows the model to effectively handle diverse and complex food images, improving its generalization capabilities.

The multi-head self-attention mechanism in Swin Transformer can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q (queries), K (keys), and V (values) are the input feature maps; d_k is the dimension of the keys.

The shifted window mechanism allows the model to capture long-range dependencies across windows, represented as:

$$\text{ShiftedWindow}(X) = \text{Roll}(X, \text{shift}) \tag{8}$$

where X is the input feature map; shift is the amount of window shifting.

The computation of window-based self-attention within a window is given by:

$$\text{WindowAttention}(Q, K, V) = \text{softmax}\left(\frac{QW^T}{\sqrt{d_k}}\right)V \tag{9}$$

where Q, K, and V are the queries, keys, and values within a specific window; W is the weight matrix.

To aggregate information across windows, a merging operation is performed, represented as:

$$\begin{aligned}\text{Merging}(X) = {} & \text{Concat}(\text{ShiftedWindow}(X_1), \text{ShiftedWindow}(X_2), \\ & \ldots, \text{ShiftedWindow}(X_n))\end{aligned} \tag{10}$$

where $X_1, X_2, \ldots, X_n$ are the feature maps from different windows; Concat represents the concatenation operation.

Finally, a feed-forward neural network (FFN) is applied to the aggregated features, represented as:

$$\text{FFN}(X) = \text{GELU}(\text{Linear}(XW_1))W_2 \tag{11}$$

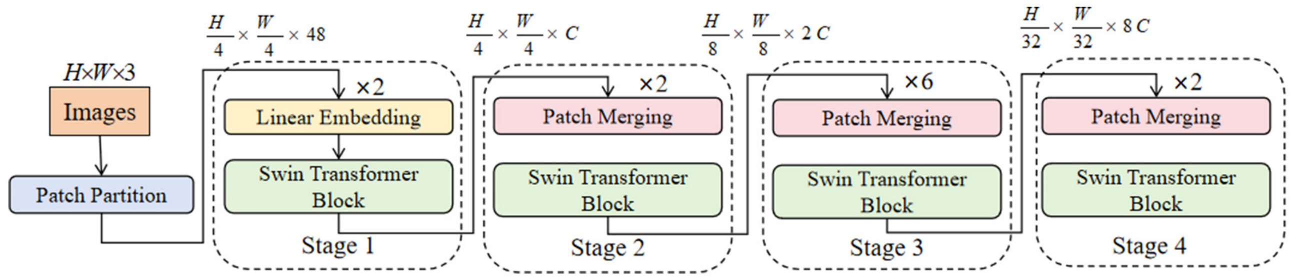


FIGURE 3
Comprehensive structure of the Swin Transformer Module (43). Adapted from (43).

where X is the aggregated feature map; W_1 and W_2 are the weights of the linear layers; GELU is the Gaussian Error Linear Unit activation function.

Additionally, the shifted window mechanism improves the model's robustness to varying lighting conditions and complex backgrounds by facilitating cross-window information exchange, thereby enhancing the stability of feature representation. Combining EfficientNet's efficient feature extraction capabilities, Swin Transformer further improves the quality of feature representation, enabling our model to classify and detect food nutrients more accurately. This approach not only enhances the accuracy of nutrient detection but also significantly reduces detection time, providing reliable technical support for food quality assessment and healthy diet monitoring. It demonstrates the immense potential and broad prospects of deep learning technology in practical applications. With the introduction of Swin Transformer, our model can more efficiently and accurately perform feature extraction and classification when processing high-resolution food images, achieving rapid and non-destructive food nutrient detection. This provides a solid technical foundation and innovative aspects for the theme of this research.

3.4 Feature Pyramid Network

In our model, we choose to use FPN to effectively fuse the multi-scale features extracted by the EfficientNet and Swin Transformer modules. FPN aims to enhance feature representation by combining high-resolution, low-level features with low-resolution, high-level features, thereby providing rich and detailed feature maps for subsequent tasks (44). FPN constructs a top-down architecture with lateral connections to fuse features at different scales. The core idea is to utilize the inherent multi-scale pyramid hierarchy of deep convolutional networks to generate feature maps at multiple levels and provide strong semantic information at all scales. This allows the model to retain high-resolution spatial information while incorporating deeper contextual and semantic information. Specifically, FPN enhances the model's ability to detect and recognize objects at various scales by combining both high-level semantic information and low-level detailed features. This multi-scale feature representation improves classification accuracy and robustness, making the model more effective in handling diverse and complex patterns in food images. The key components of FPN include the bottom-up pathway, the top-down pathway, and lateral connections. The bottom-up pathway consists of a convolutional network that extracts features at different scales. The top-down pathway is used to upsample high-level features, which are then combined with corresponding low-level features from the bottom-up pathway through lateral connections. This multi-scale feature fusion forms a feature pyramid with enhanced representational capacity. Figure 4 illustrates the structure and working principles of the FPN model.

In our model, the FPN is used to fuse the features extracted by EfficientNet and Swin Transformer, capturing fine-grained details and high-level semantic information. The FPN structure is better equipped to handle different scales and complex patterns in food images, which is crucial for accurate nutrient detection. The bottom-up pathway starts with the features extracted by EfficientNet, which contain low-level details. These features then acquire high-level semantic information through the Swin Transformer layers. In the top-down pathway, high-level features are upsampled and fused with the corresponding low-level features from EfficientNet through lateral connections. This multi-scale fusion process, which retains spatial details and incorporates semantic information, significantly enhances feature representation. By leveraging FPN, our model can generate more accurate and detailed feature maps, contributing to improved classification and detection performance. The FPN's ability to perform multi-scale feature fusion also enhances interpretability by providing better feature representations at different scales, making the model's decision-making process more transparent. The improved feature maps contribute to more accurate classification and detection of food nutrients, thereby improving the overall performance of the model.

The specific process of FPN in our model is as follows: first, initial features are extracted by EfficientNet and Swin Transformer. This process can be represented as:

$$F_l = \text{EfficientNet}(I), \quad F_h = \text{SwinTransformer}(F_l) \tag{12}$$

where I is the input image; F_l are the low-level features from EfficientNet; F_h are the high-level features from Swin Transformer.

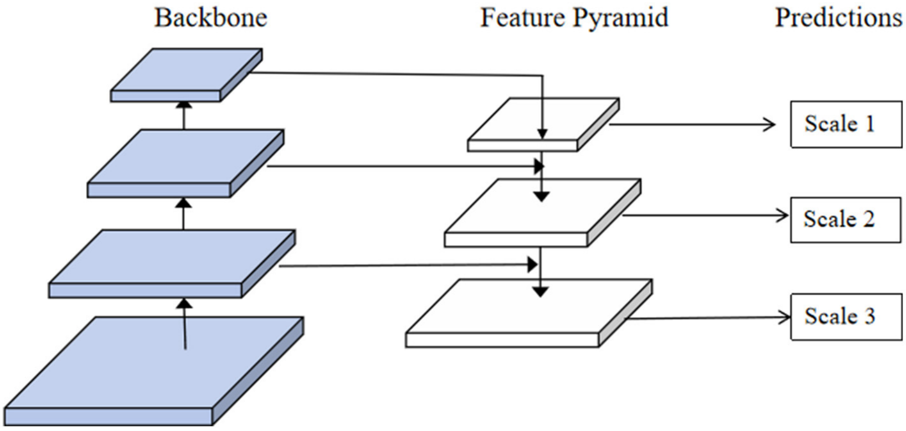


FIGURE 4
The structure of the FPN model.

Next, the top-down pathway upsamples the high-level features:

$$F_h^{\uparrow} = \text{Upsample}(F_h) \tag{13}$$

where $F_h^{\uparrow}$ denotes the upsampled high-level features.

Then, the lateral connections combine the upsampled high-level features with the low-level features:

$$F_{fusion} = F_l + F_h^{\uparrow} \tag{14}$$

where F_{fusion} is the fused feature map; + denotes element-wise addition.

The final fused feature maps at different levels are given by:

$$P_i = \text{Conv}(F_{fusion_i}) \tag{15}$$

where P_i represents the final feature maps at different levels.

By employing these operations, FPN efficiently fuses multi-scale features, providing the model with rich feature representations, thereby enhancing the accuracy and efficiency of food nutrient classification and detection. This method not only improves the accuracy of nutrient detection but also significantly reduces the detection time, offering a reliable technical support for food quality assessment and healthy diet monitoring. The integration of FPN demonstrates the significant potential and broad prospects of deep learning technology in practical applications.

3.5 Multi-task loss function

In our model, we adopt a multi-task loss function to balance the performance of four sub-tasks: calories, fat, carbohydrates, and protein. This multi-task loss function ensures that the model optimizes all these tasks simultaneously during training, thereby enhancing overall detection accuracy and robustness. The principle of the multi-task loss function is to combine the loss functions of each sub-task into a comprehensive loss function, allowing the model to learn multiple tasks simultaneously. By assigning appropriate weights to the losses of each task, the multi-task loss function ensures that the model does not overly favor any single task. This balance is crucial for achieving consistent performance across all tasks. In our model, the multi-task loss function plays a critical role, significantly improving the accuracy and efficiency of nutrient detection. By jointly optimizing multiple nutrient detection tasks, the model can leverage shared representations and the interdependencies between tasks to enhance performance. The multi-task loss function enables the model to effectively handle the complexity and diversity of food images, resulting in more accurate nutrient classification and detection.

The formula for the multi-task loss function in our model is as follows:

$$L_{\text{total}} = \sum_{t \in \{cal,protein,carb,fat\}} \left(\frac{1}{2c_t^2} L_t + \ln(1 + c_t^2) \right) \tag{16}$$

where c_t is the weight assigned to the loss of task t; L_t is the individual loss for task t.

The individual loss for each sub-task L_t is calculated using Mean Absolute Error (MAE) as follows:

$$L_t = \frac{\sum_{i=1}^{N} |y_{t,i} - \hat{y}_{t,i}|}{\sum_{i=1}^{N} y_{t,i}} \tag{17}$$

where $y_{t,i}$ is the ground truth value for the i-th sample in task t; $\hat{y}_{t,i}$ is the predicted value for the i-th sample in task t; N is the number of samples.

By combining these individual loss functions into a comprehensive multi-task loss function, our model can optimize all tasks simultaneously, thereby improving the accuracy and efficiency of food nutrient classification and detection. This approach not only enhances the prediction accuracy of the model but also improves its robustness and reliability.

4 Experiment

4.1 Experimental environment

The experiments were conducted on a high-performance workstation with the following hardware configuration: an NVIDIA Tesla V100 GPU for accelerating the training and inference processes of deep learning models; an Intel Xeon CPU providing robust computational power; and 128 GB of RAM ensuring efficient operation when handling large datasets. In terms of software environment, the experiments ran on the Ubuntu 18.04 operating system, using Python 3.8 as the programming language and PyTorch 1.8.1 as the deep learning framework to fully leverage GPU acceleration and flexible model building and training capabilities.

4.2 Datasets

This study utilizes publicly available food image datasets, including Nutrition5k (45) and ChinaMartFood109 (46). Both datasets contain various types of food images, each annotated with corresponding nutritional information. The specific details of the datasets are as follows:

4.2.1 Nutrition5k dataset

This dataset originates from a publicly available food image dataset, containing 5,000 images annotated with detailed nutritional information. The dataset comprises 5,000 static images in JPEG or PNG format, with the resolution uniformly scaled to 224 × 224 pixels. The Nutrition5k dataset is highly diverse, covering a wide range of food types, including vegetables, fruits, meats, and grains. The images were taken under different lighting conditions and backgrounds, adding to the dataset's diversity. The Nutrition5k dataset includes both 360-degree images and top-view images of food items. For the purposes of our research, we specifically utilized the top-view images from the Nutrition5k dataset. These images provide a consistent perspective that simplifies the recognition and segmentation tasks, allowing for more accurate and reliable model performance. The visualizations provided in this article correspond to the top-view images from the Nutrition5k dataset. Each visualization represents the model's output on the same image used during training and testing.

4.2.2 ChinaMartFood109 dataset

This dataset includes images of 109 categories of Chinese foods, sourced from multiple public data sources, and provides detailed nutritional annotations. The dataset comprises over 100,000 static images in JPEG or PNG format, with the resolution uniformly scaled to 224 × 224 pixels. The ChinaMartFood109 dataset covers a rich variety of Chinese food categories, including staples, dishes, snacks, and soups. The images were taken in various settings and conditions, such as restaurants, home kitchens, and outdoors, contributing to its high diversity.

By providing detailed descriptions of the datasets used, including specific parts of the Nutrition5k dataset and the conditions under which the ChinaMartFood109 images were taken, we aim to ensure the transparency and reproducibility of our research results.

4.3 Experimental details

Step1: Data preprocessing

In data preprocessing, we performed the following four key steps:

- Data cleaning: in this process, we inspected and removed images with missing values or incorrect annotations to ensure the correctness of data labels. We then used image quality detection algorithms to automatically identify and remove low-quality or blurry images. These algorithms effectively detect issues such as blurriness, noise, and resolution problems in images. A total of 350 images were identified as low-quality or blurry using these algorithms and were subsequently removed. Finally, we standardized the format and naming conventions of the images, ensuring that all images were in JPEG or PNG format for subsequent processing and analysis.
- Data standardization: we adjusted the resolution of all images to 224 × 224 pixels to meet the input requirements of the model. This step ensured that all images had consistent input sizes, which improved the training efficiency and effectiveness of the model. Additionally, we normalized the images by scaling the pixel values to a range between 0 and 1. This normalization enhanced the training effect and convergence speed of the model, preventing large numerical differences from affecting the model training.
- Data augmentation: by using techniques such as random cropping, rotation, and flipping, we augmented the images to increase data diversity and simulate different shooting angles and conditions. Furthermore, we randomly adjusted brightness, contrast, and hue to simulate various lighting conditions, thereby enhancing the model's robustness in different real-world scenarios.
- Data splitting: we divided the dataset into training, validation, and test sets in an 8:1:1 ratio. This split ratio ensures that the model has sufficient data for learning and evaluation during the training, validation, and testing phases. We also ensured consistent data distribution within each split to maintain class balance, avoiding data bias that could affect model training. These preprocessing steps ensured the quality and diversity of the data, providing a reliable foundation for subsequent model training and evaluation.

Step2: Model training

- Network parameter settings: we used EfficientNet and Swin Transformer as feature extractors and built our model on this basis. To optimize the model's performance, we chose the Adam optimizer with an initial learning rate set to 0.001 and a batch size of 32. During training, the learning rate was dynamically adjusted based on the performance of the validation set, halving every 10 epochs. The model was trained for a total of 100 epochs, and after each epoch, it was evaluated on the validation set to determine the model's convergence and stability.

- Model architecture design: our model combines the strengths of EfficientNet and Swin Transformer with specific designs based on these foundations. The EfficientNet part consists of 16 convolutional layers, with gradually increasing numbers of channels and convolution kernel sizes to better extract multi-scale features from the images. The Swin Transformer part consists of four Transformer layers, each containing 12 attention heads to capture long-range dependencies in the images. Additionally, we introduced FPN between EfficientNet and Swin Transformer, consisting of four levels of feature fusion modules. Each module includes upsampling and lateral connection operations to enhance multi-scale feature fusion. The final classification head comprises two fully connected layers with 1,024 and 512 neurons, respectively, followed by a Softmax layer for outputting classification results.
- Model training and validation process: in this study, we employed a detailed training and validation process to ensure the accuracy of the UCL model. The dataset was divided into three parts: 70% was used as the training set for training the model, 15% as the validation set for parameter tuning and early stopping control, and another 15% as the test set for evaluating the model's final performance. Additionally, we implemented five-fold cross-validation to assess the model's generalizability and robustness. Through this approach, each data subset takes turns serving as the validation set, ensuring that the model exhibits stable performance under various data conditions, thus enhancing the overall prediction accuracy and reliability. This process helps in meticulously optimizing the model to ensure its effectiveness in practical applications.

Step3: Model evaluation

In the model evaluation process, we adopted a series of evaluation metrics and cross-validation methods to comprehensively measure the model's performance in food classification and nutrient estimation tasks.

- Model performance metrics: we used multiple evaluation metrics to measure the model's performance. For the food classification task, we employed top-1 and top-5 classification accuracy. Top-1 accuracy measures the proportion of samples where the predicted class with the highest probability matches the true class, while top-5 accuracy measures the proportion of samples where the true class is among the top five predicted classes with the highest probabilities. Therefore, top-5 accuracy is typically higher than top-1 accuracy. For the nutrient estimation task, we used four evaluation metrics: Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Additionally, we calculated the 95% confidence intervals for the reported accuracies and MAPE values to ensure the statistical significance and reliability of our results.

  Mean Absolute Error (MAE) measures the average absolute difference between the predicted and ground truth values. The formula for MAE is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|n_i - \hat{n}_i| \tag{18}$$

  Mean Absolute Percentage Error (MAPE) measures the average absolute percentage difference between the predicted and ground truth values. The formula for MAPE is:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{n_i - \hat{n}_i}{n_i}\right| \times 100\% \tag{19}$$

  where N is the number of samples, n_i is the ground truth value of the i-th sample, $\hat{n}_i$ is the predicted value of the i-th sample, and $\bar{n}$ is the mean of the ground truth values.
- Cross-validation: to ensure the robustness and generalization ability of the model, we adopted cross-validation methods. Specifically, we used k-fold cross-validation (k = 5) to evaluate the model's performance. The dataset was divided into k subsets, with each iteration using k-1 subsets for training and the remaining subset for validation. This process was repeated k times, ensuring each subset was used as the validation set once. By calculating the mean and standard deviation of each validation, we obtained the model's performance across different data splits. This method effectively prevents overfitting and provides a comprehensive assessment of the model's generalization ability on different datasets. The cross-validation results indicated that our model exhibited good robustness and consistency in both food classification and nutrient estimation tasks.

  Through these evaluation methods, we can comprehensively measure the model's performance, ensuring its reliability and accuracy in practical applications. These evaluation results not only help us understand the strengths and weaknesses of the model but also provide important reference points for further optimization and improvement.

4.4 Experimental setup

To evaluate the performance of our proposed method, we conducted experiments using a diverse set of established and widely-recognized deep learning models. The selection of these seven models (VGG16, WISeR-50, Inception-V3, ResNet-152, CNN, Faster R-CNN, and Ours) was made to provide a comprehensive comparison of our method against different approaches in the field of image recognition and classification. VGG16, Inception-V3, and ResNet-152 are well-known for their performance on image classification tasks and serve as benchmarks in the field. WISeR-50 and Faster R-CNN represent more recent advancements in network architecture and object detection, respectively. Including these models allows us to evaluate our method's effectiveness across different types of architectures and highlight its advantages.

Additionally, we conducted ablation experiments to assess the contribution of different components of our proposed model. The ablation experiments involved selectively removing or altering components such as EfficientNet, Swin Transformer, and the FPN to understand their individual impact on the model's performance. This detailed analysis helps in highlighting the significance of each component in achieving the overall performance improvements.

5 Results and discussion

5.1 Comparison with existing methods

Table 1 presents the performance of our proposed method on the Nutrition5k and ChinaMartFood109 datasets, compared with several state-of-the-art models, including VGG16, WISeR50, InceptionV3, ResNet152, CNN, Faster R-CNN, DeepFood, the method by VijayaKumari et al., the method by Desai et al., and Swin-Nutrition. The results demonstrate that our method outperforms all other compared models on both datasets. On the Nutrition5k dataset, our method achieved a Top-1 accuracy of 79.50% and a Top-5 accuracy of 95.66%, while the best-performing existing model, InceptionV3, achieved a Top-1 accuracy of 72.13% and a Top-5 accuracy of 93.42%. Additionally, the DeepFood model achieved a Top-1 accuracy of 74.10% and a Top-5 accuracy of 92.20%, and the Swin-Nutrition model achieved a Top-1 accuracy of 70.45% and a Top-5 accuracy of 89.55%. These results indicate that our method offers significant advantages in food image classification tasks. On the ChinaMartFood109 dataset, our method achieved a Top-1 accuracy of 80.25% and a Top-5 accuracy of 96.98%. In comparison, the best-performing existing model, InceptionV3, achieved a Top-1 accuracy of 78.26% and a Top-5 accuracy of 96.62%. Additionally, the DeepFood model achieved a Top-1 accuracy of 73.70% and a Top-5 accuracy of 91.80%, and the Swin-Nutrition model achieved a Top-1 accuracy of 69.30% and a Top-5 accuracy of 89.10%. Although the improvements are relatively modest, these results still highlight the advantages of our method in handling large-scale, multi-category food datasets. By integrating EfficientNet and Swin Transformer, our method can better handle the diversity and complexity of food images. Additionally, the FPN enhances feature representation through multi-scale feature fusion, further improving classification accuracy and robustness. The experimental results demonstrate that our method significantly outperforms existing methods in the field of food nutrient recognition.

It is worth noting that although the Swin-Nutrition model is somewhat similar to our model, it does not perform as well in terms of accuracy and robustness. Swin-Nutrition primarily relies on Swin Transformer as the backbone network for feature extraction and uses a Feature Fusion Module (FFM) and a nutrient prediction module for evaluation. While this method shows some effectiveness on the Nutrition5k dataset, its results do not surpass those of our proposed method. Compared to Swin-Nutrition, our method improves and optimizes feature extraction and fusion by adopting EfficientNet as the base feature extraction network and combining it with Swin Transformer to further enhance feature representation. Our model also incorporates FPN to improve robustness and accuracy through multi-scale feature fusion. The experimental results indicate that our method performs exceptionally well across different datasets, demonstrating better generalizability and adaptability. In conclusion, our method shows higher accuracy and robustness in food image classification and nutrient content estimation tasks, significantly outperforming the existing Swin-Nutrition model and other compared methods.

TABLE 1 Comparison of state-of-the-art methods on Nutrition5k and ChinaMartFood109 datasets.

Dataset	Model	Top-1%	Top-5%
Nutrition5k	VGG16	54.21	77.61
	WISeR50	68.32	89.75
	InceptionV3	72.13	93.42
	ResNet152	75.64	94.55
	CNN	58.27	81.92
	Faster R-CNN	65.78	90.23
	DeepFood (47)	74.10	92.20
	VijayaKumari et al. (48)	68.50	88.30
	Desai et al. (49)	71.20	90.00
	Swin-Nutrition (25)	70.45	89.55
	Ours	79.50	95.66
ChinaMartFood109	VGG16	52.17	79.65
	WISeR50	77.16	95.21
	InceptionV3	78.26	96.62
	ResNet152	76.53	92.09
	CNN	55.21	82.17
	Faster R-CNN	69.58	91.13
	DeepFood (47)	73.70	91.80
	VijayaKumari et al. (48)	67.20	87.50
	Desai et al. (49)	70.90	89.70
	Swin-Nutrition (25)	69.30	89.10
	Ours	80.25	96.98

Figure 5 shows the top-1 and top-5 accuracy for each food class and category group. In Figure 5A, the variation in image recognition accuracy among different food classes is evident. While some classes achieve high accuracy, others show significant variability, indicating that certain food items are more challenging for the model to predict accurately. Figure 5B illustrates the differences in recognition performance across various food category groups. Categories such as "Braised Beef with Brown Sauce" and "Tomato and Egg Soup" consistently show high accuracy, suggesting the model's strong performance in these groups. Conversely, categories like "Minced Sauteed Celery" and "Braised Pork Leg" exhibit lower accuracy, reflecting the model's difficulty in recognizing these food types accurately. The results suggest that the model performs better in more common and well-represented food categories. However, the variability in accuracy indicates the need for further model refinement and the inclusion of

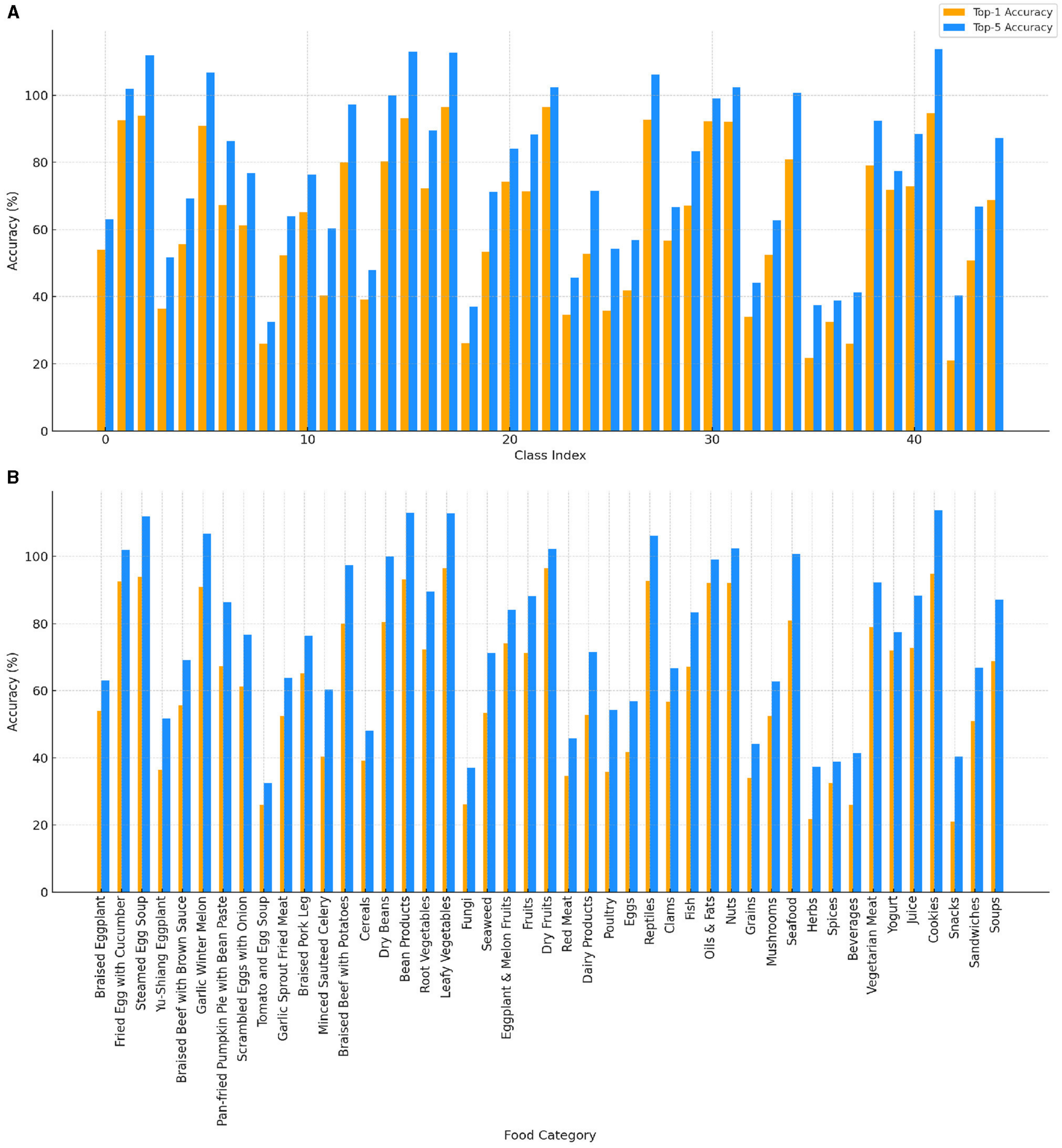


FIGURE 5
Top-1 (orange) and top-5 (dodger blue) accuracy for each class and category group. **(A)** Accuracy across different class indices, showing variation in image recognition accuracy. **(B)** Accuracy for each food category, highlighting differences in recognition performance across various food groups.

more diverse training data. This analysis provides a clear direction for refining the model and improving its overall performance in food classification and nutrient recognition tasks.

From the data in Table 2, it is evident that our method outperforms existing approaches across different datasets and evaluation metrics. On the Nutrition5k dataset, our method achieves the lowest Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for calorie prediction, with values of 37.90 and 14.72%, respectively. In comparison, VGG16 records a calorie MAE and MAPE of 54.21 and 18.82%, and WISeR50 records 50.32 and 16.52%, indicating a significant performance improvement in calorie prediction by our method. Similarly, our method shows excellent performance for fat, carbohydrate, and protein predictions, with MAE and MAPE

TABLE 2 Comparison of the performance of different methods.

Dataset	Methods	Calorie MAE/MAPE	Fat MAE/MAPE	Carb MAE/MAPE	Protein MAE/MAPE	Mean MAPE
Nutrition5k	VGG16	54.21/18.82%	2.27/18.12%	4.60/23.84%	3.70/20.91%	20.17%
	WISeR50	50.32/16.52%	2.50/21.03%	4.20/22.01%	3.40/19.54%	19.77%
	InceptionV3	47.13/15.32%	2.90/20.64%	4.10/21.49%	3.20/18.87%	18.58%
	ResNet152	49.64/16.23%	3.10/22.47%	4.32/22.21%	3.80/20.49%	20.35%
	CNN	58.27/26.11%	5.00/34.20%	6.10/31.92%	5.50/29.49%	29.10%
	Faster R-CNN	55.78/19.23%	3.00/20.12%	4.50/23.03%	3.60/21.04%	20.85%
	DeepFood (47)	44.10/15.12%	2.70/18.44%	4.00/20.64%	3.30/18.31%	18.63%
	VijayaKumari et al. (48)	50.50/16.82%	2.80/19.34%	4.40/21.04%	3.50/19.42%	19.65%
	Desai et al. (49)	46.90/15.72%	2.75/18.74%	3.95/20.54%	3.25/18.25%	18.81%
	Swin-Nutrition (25)	45.40/15.3%	2.60/22.1%	4.05/20.8%	3.20/15.4%	18.4%
	Ours	37.90/14.72%	2.60/18.04%	3.90/19.84%	3.10/18.02%	17.80%
ChinaMartFood109	VGG16	62.17/20.52%	3.10/22.44%	5.20/24.49%	4.00/23.12%	22.57%
	WISeR50	58.16/19.03%	3.00/21.23%	5.00/23.01%	3.80/21.54%	21.20%
	InceptionV3	60.26/19.79%	3.20/22.04%	5.10/23.41%	3.90/22.02%	21.82%
	ResNet152	59.53/19.62%	3.30/22.76%	5.20/24.19%	4.10/23.32%	22.47%
	CNN	66.21/27.11%	5.50/34.78%	6.50/31.98%	5.90/30.51%	31.10%
	Faster R-CNN	63.58/20.23%	3.20/22.01%	5.00/23.50%	4.00/22.12%	22.00%
	DeepFood (47)	51.70/18.92%	2.90/19.54%	4.60/21.84%	3.70/20.01%	20.08%
	VijayaKumari et al. (48)	59.10/19.33%	3.20/21.74%	5.10/23.24%	3.90/21.92%	21.56%
	Desai et al. (49)	53.90/19.02%	3.00/20.84%	4.75/22.14%	3.80/20.72%	20.68%
	Swin-Nutrition (25)	52.30/18.12%	2.90/20.14%	4.60/22.04%	3.80/19.22%	19.88%
	Ours	41.90/15.21%	2.80/18.54%	4.30/20.04%	3.50/18.91%	18.18%

metrics outperforming other models. In the ChinaMartFood109 dataset, our method also demonstrates superior performance in predicting calories, fat, carbohydrates, and protein, as reflected by the MAE and MAPE metrics. For instance, in calorie prediction, our method achieves an MAE and MAPE of 41.90 and 15.21%, respectively, significantly lower than VGG16's 62.17 and 20.52%, and ResNet152's 59.53 and 19.62%. This outstanding performance underscores the stability and robustness of our method across different datasets. Furthermore, the data reveal that our method achieves the lowest average MAPE (Mean MAPE) across all evaluation metrics on both datasets, with values of 17.80 and 18.18 for the Nutrition5k and ChinaMartFood109 datasets, respectively. In contrast, traditional CNN models exhibit higher average MAPEs of 29.10 and 31.10 on the Nutrition5k and ChinaMartFood109 datasets, respectively. This further validates the advantage of our method in multi-task nutritional content detection. Notably, in the Nutrition5k dataset, while our method performs excellently overall, Swin-Nutrition shows outstanding performance in certain metrics, such as a MAPE of 15.4% for protein. This phenomenon may be due to Swin-Nutrition using the Swin Transformer for feature extraction, which excels in global feature extraction when processing certain types of food images. However, considering all metrics, our method still outperforms Swin-Nutrition overall, especially in predicting fat and carbohydrates, where our method demonstrates superior performance. These results indicate that by leveraging the efficient feature extraction and deep feature fusion capabilities of EfficientNet and Swin Transformer, our method can more accurately predict the nutritional content of food items. This not only enhances detection accuracy and efficiency but also demonstrates the significant potential and broad applicability of our method in real-world applications.

Our model's ability to address issues related to data processing, model robustness, and interpretability is clearly demonstrated in these results. The combination of EfficientNet and Swin Transformer allows for effective handling of varying lighting conditions and complex food compositions, enhancing model robustness. Additionally, the FPN enhances multi-scale feature fusion, improving the interpretability of the model by providing better feature representations at different scales. These innovations simplify the data processing workflow and enhance the model's generalization ability, making it a more robust and accurate solution for food nutrient detection compared to existing methods.

Figure 6 illustrate the relationship between predicted and actual values for seven models (VGG16, WISeR-50, Inception-V3, ResNet-152, CNN, Faster R-CNN, and Ours) across four nutritional components (calories, fat, carbohydrates, and protein). Each plot includes scatter points, a diagonal line, and confidence ellipses. Each scatter point represents the relationship between a

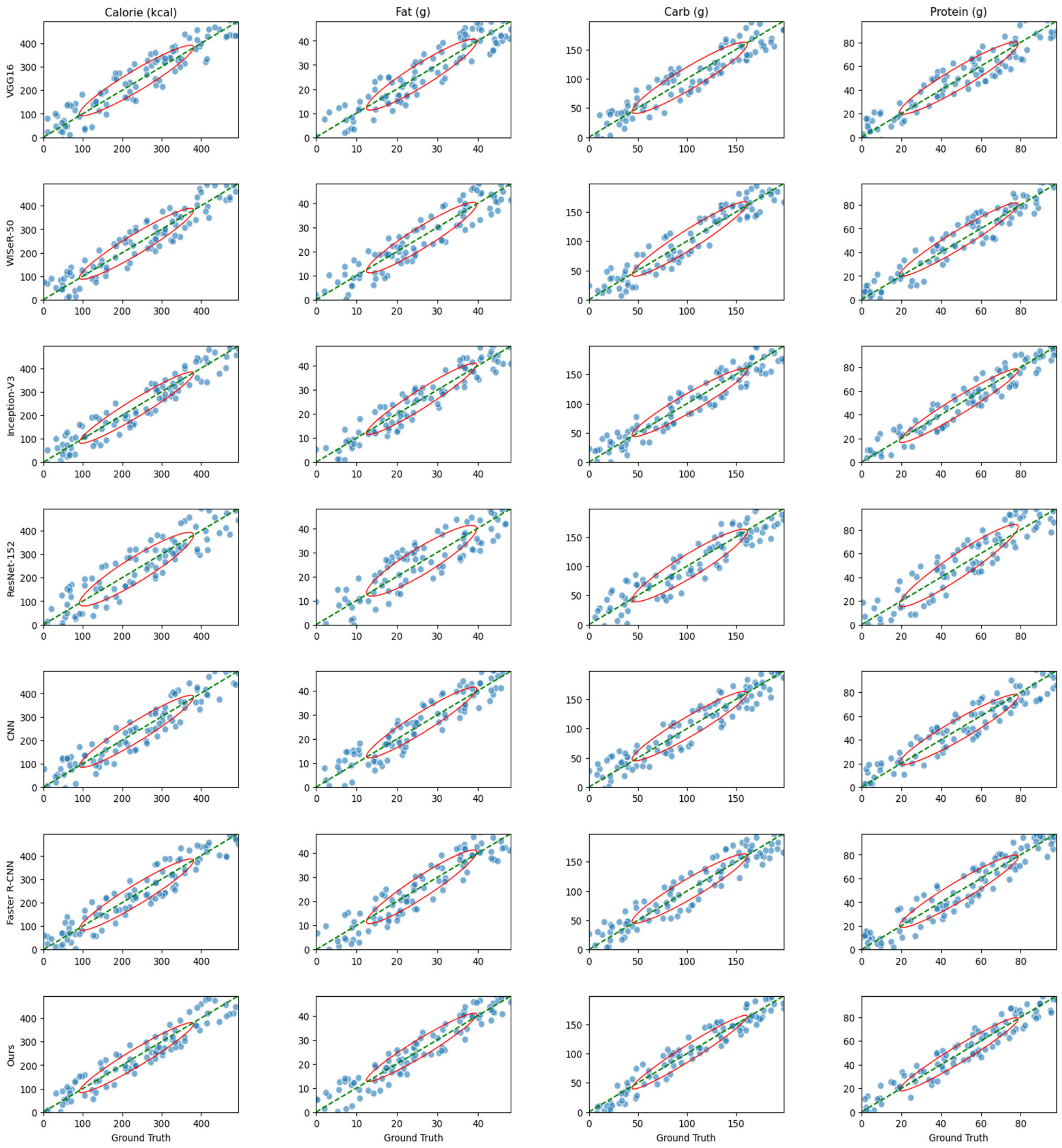


FIGURE 6
Scatter plots comparing the predicted and actual values of four nutritional components (calories, fat, carbohydrates, and protein) for seven models (VGG16, WISeR-50, Inception-V3, ResNet-152, CNN, Faster R-CNN, and Ours).

model's predicted value and the actual value for a food sample. The diagonal line indicates the ideal scenario where the predicted values perfectly match the actual values. Most scatter points are clustered around the diagonal line, indicating overall good predictive performance of the models. However, some points deviate from the diagonal, suggesting that certain models exhibit prediction errors in specific cases. The confidence ellipses show the distribution range of the data points, with the size and shape reflecting the variance and covariance of the predicted values. Smaller ellipses indicate lower variance and more stable predictions. These ellipses allow for a visual comparison of the prediction distributions among different models. From the plots, it is evident that VGG16 has relatively larger confidence ellipses for all nutritional components, indicating higher variance and

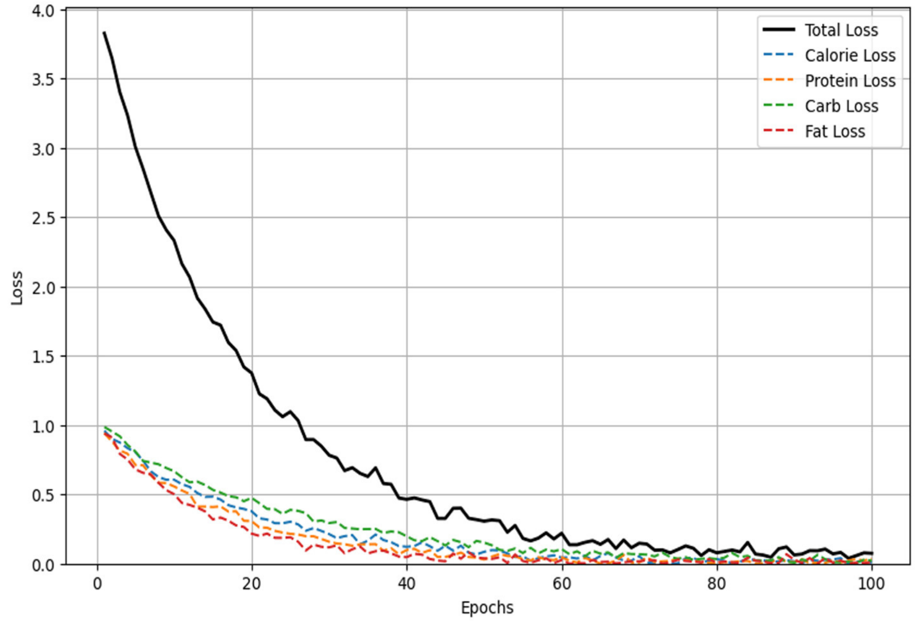


FIGURE 7
Multimodal loss curves over 100 epochs for the prediction of four nutritional components (calories, protein, carbohydrates, and fat) and the total loss.

less stable predictions. In contrast, WISeR-50 and Inception-V3 show smaller ellipses, indicating more concentrated and stable predictions. ResNet-152 and Faster R-CNN also display relatively small ellipses, demonstrating good predictive performance. CNN, however, has larger ellipses, indicating higher variance and poorer performance compared to other models. Overall, the Ours model exhibits the smallest confidence ellipses across all nutritional components, indicating the least variance and highest stability and accuracy in predictions. This demonstrates the significant advantage of the Ours model in predicting food nutritional components. In summary, the plots reveal the performance differences among the models in predicting various nutritional components. VGG16 and CNN exhibit more dispersed predictions with higher variance, whereas WISeR-50, Inception-V3, ResNet-152, Faster R-CNN, and Ours show more concentrated predictions with better performance. Notably, the Ours model outperforms all other models, demonstrating superior predictive performance and stability.

5.2 Multimodal loss analysis

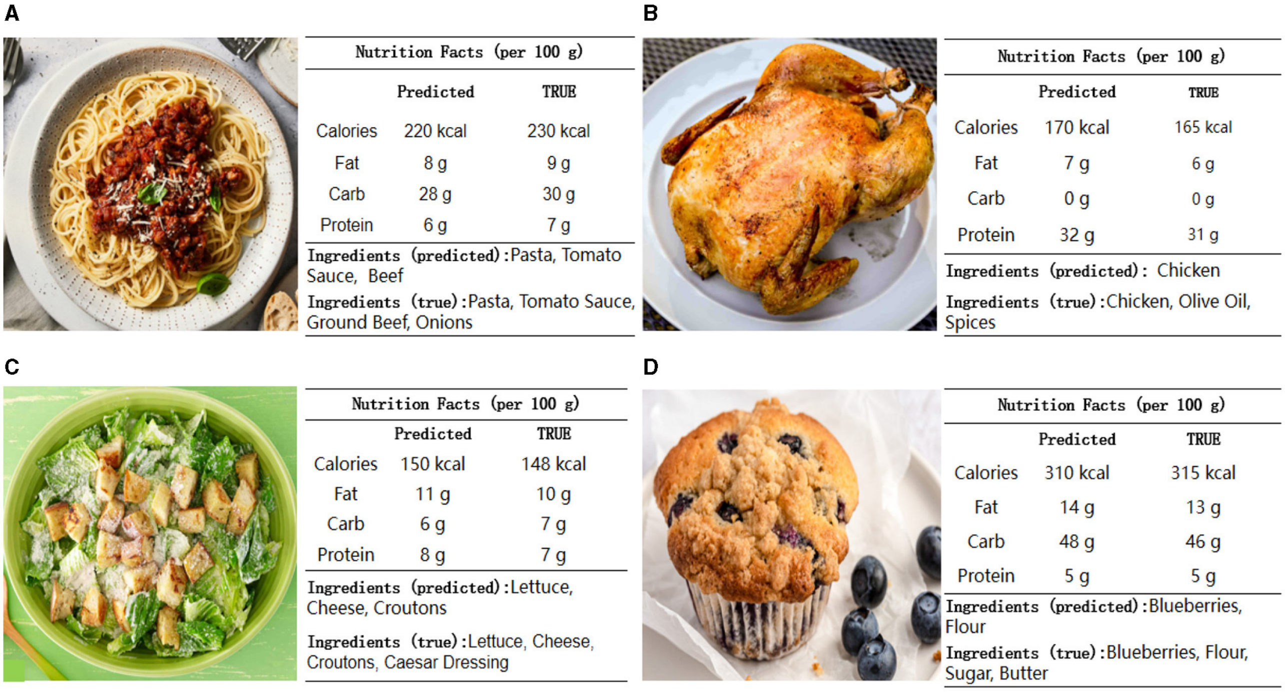
Figure 7 illustrates the changes in the multimodal loss over 100 training epochs for four nutritional components (calories, protein, carbohydrates, and fat) and the total loss. As the training epochs progress, both the total loss (represented by the solid black line) and the individual losses (represented by dashed lines) exhibit a significant downward trend, indicating an improvement in the model's predictive accuracy for each sub-task. The loss for fat decreases the fastest, demonstrating the model's superior learning efficiency for this particular task, while the losses for calories, protein, and carbohydrates also decrease steadily, reflecting effective optimization in these areas as well. The overall decline in total loss highlights the enhancement of the model's performance across all tasks. The multimodal loss function effectively balances the losses of each sub-task, ensuring that the model optimizes all tasks simultaneously during the training process. This balanced approach prevents the model from overfitting to any single task, thus achieving consistent performance across all tasks. Furthermore, all loss curves show a clear convergence trend, especially within the first 50 epochs where the losses decrease most rapidly before leveling off, indicating that the model is approaching a stable state. This suggests that the model quickly learns effective features in the initial training phase, with further fine-tuning occurring as training progresses. In summary, these loss curves validate the effectiveness of the multimodal loss function in enhancing predictive accuracy and robustness. By jointly optimizing multiple nutritional component prediction tasks, the model leverages shared representations and inter-task dependencies to improve overall performance. The results demonstrate that the model, utilizing a multimodal loss function, achieves high predictive accuracy and stability when handling the complex task of food nutritional component prediction.

5.3 Ablation experiment

Table 3 presents the results of ablation experiments on the Nutrition5k and ChinaMartFood109 datasets, specifically examining the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for calorie, fat, carbohydrate, and protein predictions. The results indicate that different model configurations have a significant impact on performance. For the Nutrition5k dataset, the configuration using only EfficientNet shows relatively high errors, with a calorie MAPE of 21.15%. When using only the Swin Transformer, the calorie MAPE decreases to 19.82%,

TABLE 3 Ablation study results on Nutrition5k and ChinaMartFood109 datasets.

Dataset	EfficientNet	Swin transformer	FPN	Calorie MAE/MAPE	Fat MAE/MAPE	Carb MAE/MAPE	Protein MAE/MAPE	Mean MAPE
Nutrition5k	✓	×	×	45.32/21.15%	4.11/23.45%	6.32/27.34%	5.23/24.21%	24.53%
	×	✓	×	42.50/19.82%	3.80/22.34%	5.80/25.45%	4.80/22.11%	22.93%
	✓	✓	×	38.44/17.89%	3.22/19.56%	5.11/23.12%	4.01/21.13%	20.93%
	✓	✓	✓	37.90/14.72%	2.60/18.04%	3.90/19.84%	3.10/18.02%	17.80%
ChinaMartFood109	✓	×	×	58.12/24.32%	5.22/28.43%	7.45/30.54%	6.21/27.22%	27.63%
	×	✓	×	54.30/22.45%	4.90/26.12%	7.10/29.22%	5.90/26.11%	26.48%
	✓	✓	×	46.23/20.34%	4.01/24.21%	6.12/25.78%	5.02/23.45%	23.45%
	✓	✓	✓	41.90/15.21%	2.80/18.54%	4.30/20.04%	3.50/18.91%	18.18%



A

Nutrition Facts (per 100 g)	Predicted	TRUE
Calories	220 kcal	230 kcal
Fat	8 g	9 g
Carb	28 g	30 g
Protein	6 g	7 g

Ingredients (predicted):Pasta, Tomato Sauce, Beef

Ingredients (true):Pasta, Tomato Sauce, Ground Beef, Onions

B

Nutrition Facts (per 100 g)	Predicted	TRUE
Calories	170 kcal	165 kcal
Fat	7 g	6 g
Carb	0 g	0 g
Protein	32 g	31 g

Ingredients (predicted): Chicken

Ingredients (true):Chicken, Olive Oil, Spices

C

Nutrition Facts (per 100 g)	Predicted	TRUE
Calories	150 kcal	148 kcal
Fat	11 g	10 g
Carb	6 g	7 g
Protein	8 g	7 g

Ingredients (predicted):Lettuce, Cheese, Croutons

Ingredients (true):Lettuce, Cheese, Croutons, Caesar Dressing

D

Nutrition Facts (per 100 g)	Predicted	TRUE
Calories	310 kcal	315 kcal
Fat	14 g	13 g
Carb	48 g	46 g
Protein	5 g	5 g

Ingredients (predicted):Blueberries, Flour

Ingredients (true):Blueberries, Flour, Sugar, Butter

FIGURE 8
Application of the model in real-world scenarios: Comparison of predicted and true nutritional values and ingredients for various food items. The figure illustrates the accuracy and reliability of the model in predicting the nutritional content of **(A)** Spaghetti Bolognese, **(B)** Grilled Chicken, **(C)** Caesar Salad, and **(D)** Blueberry Muffin.

demonstrating Swin Transformer's advantage in capturing long-range dependencies in images. When combining EfficientNet and Swin Transformer, the calorie MAPE further reduces to 17.89%, indicating a synergistic effect in feature extraction and representation. Finally, when integrating the FPN with EfficientNet and Swin Transformer, the model achieves the best performance, with a calorie MAPE of only 14.72%, highlighting FPN's crucial role in multi-scale feature fusion. Similarly, on the ChinaMartFood109 dataset, the configuration using only EfficientNet has a calorie MAPE of 24.32%. Adding the Swin Transformer reduces the calorie MAPE to 22.45%. When combining EfficientNet and Swin Transformer, the calorie MAPE further decreases to 20.34%. The best performance is achieved when all three components are combined, resulting in a calorie MAPE of 15.21%. The consistent trend across both datasets underscores the comprehensive roles of EfficientNet's feature extraction capabilities, Swin Transformer's attention mechanisms, and FPN's feature fusion techniques in enhancing model accuracy and robustness. Overall, the ablation experiment results demonstrate that each component of our proposed model significantly contributes to the overall performance. EfficientNet provides a strong foundation for feature extraction, Swin Transformer enhances the model's ability to capture complex patterns, and FPN ensures effective multi-scale feature fusion.

TABLE 4 Examples of the nutritional composition from the Nutritions5 dataset.

Ingredient	Calorie (kCal)	Fat (g)	Carb (g)	Protein (g)
Dish_1560442303				
Berries	83.78	0.43	20.57	1.02
Scrambled eggs	111.15	8.21	1.17	7.58
Bacon	205.54	15.89	0.52	14.08
Quinoa	44.43	0.72	7.79	1.59
Sum	444.9	25.25	30.05	24.27
Dish_1562788816				
Pork	231.05	13.64	0	25.28
Fish	36.08	5.28	0	7.45
Rice	133.14	0.34	28.74	2.76
Corn	41.11	0.98	8.64	1.17
Sum	441.38	20.24	37.38	36.66
Dish_1558114609				
Grapes	26.87	0.07	7.05	0.29
Sausage	259.04	20.46	3.18	14.43
Bacon	151.44	11.75	0.38	10.39
Almonds	104.08	8.88	4.05	3.98
Sum	541.43	41.16	14.66	29.09

5.4 Visualization analysis

The visualization analysis presents a comparison between predicted and true nutritional values for four different food items: Spaghetti Bolognese, Grilled Chicken, Caesar Salad, and Blueberry Muffin as shown in Figure 8. Across these food items, the predicted values for calories, fat, carbohydrates, and protein generally align closely with the true values, demonstrating the model's accuracy and robustness. The predictions for macronutrients like calories, fat, carbohydrates, and protein show only minor discrepancies. The model accurately identifies key ingredients in simpler dishes, such as Grilled Chicken, where predictions are precise. For more complex dishes, like Spaghetti Bolognese and Caesar Salad, the model performs well but misses some specific ingredients, such as croutons and dressing in the Caesar Salad or specific types of meat in the Bolognese. Overall, the model's ability to predict nutritional values and identify ingredients is highly reliable, making it a valuable tool for nutritional analysis. The minor discrepancies observed are within acceptable ranges, affirming the model's practical applicability. Future enhancements could focus on improving ingredient identification for more complex recipes, further boosting the model's accuracy and utility in real-world scenarios.

Table 4 shows the nutritional composition of different dishes in the Nutritions5 dataset, which is used to evaluate the effectiveness of our proposed deep learning nutrition recognition model. For Dish_1560442303, the total calories are 444.9 kCal, containing 25.25 g of fat, 30.05 g of carbohydrates, and 24.27 g of protein. The main nutritional sources of this dish are bacon and scrambled eggs, which provide high levels of calories and protein, while berries and quinoa contribute a large amount of carbohydrates. Dish_156278816 has a total of 441.38 kCal, with 20.24 g of fat, 37.38 g of carbohydrates, and 36.66 g of protein. In this dish, pork and fish are the main sources of calories and protein, while rice and corn provide most of the carbohydrates. For Dish_1558114609, the total calories are 541.43 kCal, with 41.16 g of fat, 14.66 g of carbohydrates, and 29.09 g of protein. Sausage and bacon are the main sources of calories and fat, while grapes and almonds provide higher amounts of carbohydrates and protein. Figure 9 presents the visualizations of these dishes' nutritional composition, intuitively displaying the distribution of calories, fat, carbohydrates, and protein in different foods. The results in Table 4 and the visualizations in Figure 9 indicate that our proposed deep learning model can effectively recognize and classify various nutritional components in different food images. This demonstrates the robustness and adaptability of the model in accurately identifying nutritional information, which is crucial for precise dietary assessment and monitoring. The model can quickly and non-destructively detect food components, providing reliable technical support for food quality evaluation and healthy diet monitoring. In conclusion, the proposed model significantly improves the accuracy and efficiency of food nutrition recognition, providing a solid foundation for intelligent and refined analysis of nutritional components. This advancement not only supports food quality evaluation but also has wide applications and social value in promoting healthy eating habits.

6 Conclusion

In this study, we proposed a novel deep learning model that integrates EfficientNet, Swin Transformer, and Feature Pyramid Network (FPN) to enhance the accuracy and efficiency of food nutrient recognition. Our model was evaluated on two extensive datasets, Nutrition5k and ChinaMartFood109, where it demonstrated superior performance compared to several state-of-the-art models. The experimental results indicated that our model achieved higher Top-1 and Top-5 accuracy rates for food classification tasks and significantly lower Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for nutrient estimation tasks. Additionally, visualization analysis of the predicted vs. actual nutritional values for various food items showed that our model could reliably and accurately predict nutritional content, affirming its practical applicability in real-world scenarios. Furthermore, our model addresses several key issues identified in existing methods, such as data processing efficiency, model robustness, and interpretability. By integrating EfficientNet for efficient feature extraction, Swin Transformer for capturing long-range dependencies, and FPN for multi-scale feature fusion, our model provides a comprehensive approach to improve food nutrient recognition accuracy. The FPN enhances interpretability by performing multi-scale feature fusion, which not only improves classification accuracy but also provides better feature representations at different scales, making the model's decision-making process more transparent. Despite the promising results, our model has certain limitations. Firstly, it struggles

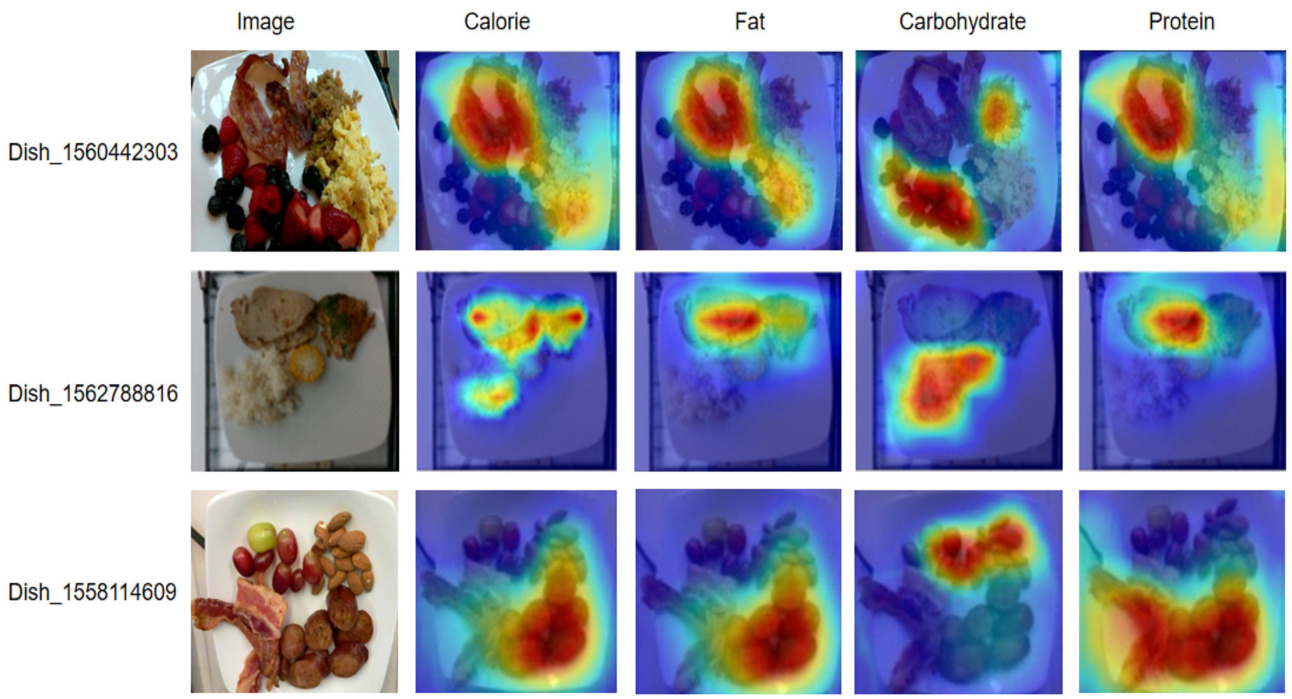


FIGURE 9
Visualizations of nutritional composition for different dishes from the Nutrition 5k dataset.

with complex dishes containing multiple ingredients, occasionally missing specific components such as croutons and certain types of meat, which could affect overall nutritional estimation. Secondly, the training process requires substantial computational resources due to the complex architecture, limiting accessibility and scalability for users with limited resources. Thirdly, the model heavily relies on high-quality images for optimal performance. In real-world scenarios, varying image quality due to different lighting conditions, occlusions, and variations in image resolution can adversely affect the model's accuracy and reliability.

Looking forward, future work should aim to address these limitations by incorporating more comprehensive and diverse datasets that include a wider range of complex dishes and their detailed ingredient annotations. Enhancing the model's ability to disaggregate and accurately recognize multiple ingredients within a single dish will improve its performance in complex food recognition tasks. Additionally, optimizing the model architecture to reduce computational requirements without compromising accuracy could make it more accessible and scalable for broader use. Developing techniques to handle varying image qualities and enhancing the model's robustness to different environmental conditions will be crucial for its real-world applicability. Advanced techniques such as transfer learning and incremental learning could also be explored to maintain high performance while reducing the computational burden.

In conclusion, our model significantly advances food nutrient recognition by combining EfficientNet, Swin Transformer, and FPN. It achieves high accuracy and robustness in nutrient estimation and is practical for real-world dietary assessment and monitoring. This work lays a foundation for future research in intelligent food analysis systems, potentially impacting health and nutrition sectors by providing reliable tools for dietary management and food quality assessment. The model's rapid and non-destructive nutrient detection enhances understanding of food nutrition and promotes healthy dietary practices.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HW: Data curation, Formal analysis, Investigation, Resources, Writing – original draft. HT: Investigation, Methodology, Project administration, Resources, Writing – review & editing. RJ: Software, Supervision, Validation, Visualization, Writing – review & editing. LM: Conceptualization, Data curation, Formal analysis, Writing – review & editing. LY: Software, Supervision, Validation, Visualization, Writing – review & editing. JC: Investigation, Project administration, Supervision, Visualization, Writing – review & editing. FL: Conceptualization, Formal analysis, Methodology, Resources, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

HT was employed by China Tea Technology Co., Ltd. FL was employed by Beijing Sanyuan Foods Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhou Y, Tian Y, Ollennu-Chuasam P, Kortesniemi M, Selander K, Väänänen K, et al. Compositional characteristics of red clover (*Trifolium pretense*) seeds and supercritical CO_2 extracted seed oil as potential sources of bioactive compounds. *Food Innov Adv*. (2024) 3:11–9. doi: 10.48130/fia-0024-0002

2. Ali A, Ali S, Husnain M, Saad Missen MM, Samad A, Khan M. Detection of deficiency of nutrients in grape leaves using deep network. *Math Probl Eng*. (2022) 2022:3114525. doi: 10.1155/2022/3114525

3. Li G, Zou H, Chen Y. Discrimination capacity analysis of FTIR-PCA and EEM-PARAFAC on dandelion tissues extracts. *Food Innov Adv*. (2023) 2:247–54. doi: 10.48130/FIA-2023-0026

4. Kumar Y, Dubey AK, Arora RR, Rocha A. Multiclass classification of nutrients deficiency of apple using deep neural network. *Neural Comput Appl*. (2022) 34:1–12. doi: 10.1007/s00521-020-05310-x

5. Taha MF, Abdalla A, ElMasry G, Gouda M, Zhou L, Zhao N, et al. Using deep convolutional neural network for image-based diagnosis of nutrient deficiencies in plants grown in aquaponics. *Chemosensors*. (2022) 10:45. doi: 10.3390/chemosensors10020045

6. Talukder MSH, Sarkar AK. Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning. *Smart Agric Technol*. (2023) 4:100155. doi: 10.1016/j.atech.2022.100155

7. Ródenas J, Nagarajan B, Bolaños M, Radeva P. Learning multi-subset of classes for fine-grained food recognition. In: *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management*. New York, NY: ACM (2022), p. 17–26. doi: 10.1145/3552484.3555754

8. Costa L, Kunwar S, Ampatzidis Y, Albrecht U. Determining leaf nutrient concentrations in citrus trees using UAV imagery and machine learning. *Precision Agric*. (2022) 23:854–75. doi: 10.1007/s11119-021-09864-1

9. Kareem RSA, Tilford T, Stoyanov S. Fine-grained food image classification and recipe extraction using a customized deep neural network and NLP. *Comput Biol Med*. (2024) 175:108528. doi: 10.1016/j.compbiomed.2024.108528

10. Zhang Y, Deng L, Zhu H, Wang W, Ren Z, Zhou Q, et al. Deep learning in food category recognition. *Inf. Fusion*. (2023) 98:101859. doi: 10.1016/j.inffus.2023.101859

11. Wang D, Gan J, Mao J, Chen F, Yu L. Forecasting power demand in China with a CNN-LSTM model including multimodal information. *Energy*. (2023) 263:126012. doi: 10.1016/j.energy.2022.126012

12. Lu Y, Chen D, Olaniyi E, Huang Y. Generative adversarial networks (GANs) for image augmentation in agriculture: a systematic review. *Comput Electron Agric*. (2022) 200:107208. doi: 10.1016/j.compag.2022.107208

13. Alahmari SS, Salem T. Food state recognition using deep learning. *IEEE Access*. (2022) 10:130048–57. doi: 10.1109/ACCESS.2022.3228701

14. Chopra M, Purwar A. Recent studies on segmentation techniques for food recognition: a survey. *Arch Computat Methods Eng*. (2022) 29:865–78. doi: 10.1007/s11831-021-09598-3

15. Chakravartula SSN, Moscetti R, Bedini G, Nardella M, Massantini R. Use of convolutional neural network (CNN) combined with FT-NIR spectroscopy to predict food adulteration: a case study on coffee. *Food Control*. (2022) 135:108816. doi: 10.1016/j.foodcont.2022.108816

16. Wang C, Liu S, Wang Y, Xiong J, Zhang Z, Zhao B, et al. Application of convolutional neural network-based detection methods in fresh fruit production: a comprehensive review. *Front Plant Sci*. (2022) 13:868745. doi: 10.3389/fpls.2022.868745

17. Shah B, Bhavsar H. Depth-restricted convolutional neural network-a model for Gujarati food image classification. *Vis Comput*. (2024) 40:1931–46. doi: 10.1007/s00371-023-02893-z

18. Wang J, Li F, An Y, Zhang X, Sun H. Towards robust LiDAR-camera fusion in BEV space via mutual deformable attention and temporal aggregation. *IEEE Trans Circ Syst Video Technol*. (2024) 34:5753–64. doi: 10.1109/TCSVT.2024.3366664

19. Liu YC, Onthoni DD, Mohapatra S, Irianti D, Sahoo PK. Deep-learning-assisted multi-dish food recognition application for dietary intake reporting. *Electronics*. (2022) 11:1626. doi: 10.3390/electronics11101626

20. Qian C, Liu H, Du T, Sun S, Liu W, Zhang R. An improved U-Net network-based quantitative analysis of melon fruit phenotypic characteristics. *J Food Meas Charact*. (2022) 16:4198–207. doi: 10.1007/s11694-022-01519-7

21. Kursun R, Bastas KK. Koklu M. Segmentation of dry bean (*Phaseolus vulgaris* L) leaf disease images with U-Net and classification using deep learning algorithms. *Eur Food Res Technol*. (2023) 249:2543–58. doi: 10.1007/s00217-023-04319-5

22. Konstantakopoulos FS, Georga EI, Fotiadis DI. A review of image-based food recognition and volume estimation artificial intelligence systems. *IEEE Rev Biomed Eng*. (2023) 17:136–52. doi: 10.1109/RBME.2023.3283149

23. Sharma U, Artacho B, Savakis A. Gourmetnet: food segmentation using multi-scale waterfall features with spatial and channel attention. *Sensors*. (2021) 21:7504. doi: 10.3390/s21227504

24. Liang H, Wen G, Hu Y, Luo M, Yang P, Xu Y. MVANet: multi-task guided multi-view attention network for Chinese food recognition. *IEEE Trans Multimed*. (2020) 23:3551–61. doi: 10.1109/TMM.2020.3028478

25. Shao W, Hou S, Jia W, Zheng Y. Rapid non-destructive analysis of food nutrient content using swin-nutrition. *Foods*. (2022) 11:3429. doi: 10.3390/foods11213429

26. Aguilar E, Bolaños M, Radeva P. Regularized uncertainty-based multi-task learning model for food analysis. *J Vis Commun Image Represent*. (2019) 60:360–70. doi: 10.1016/j.jvcir.2019.03.011

27. Min W, Wang Z, Liu Y, Luo M, Kang L, Wei X, et al. Large scale visual food recognition. *IEEE Trans Pattern Anal Mach Intell*. (2023) 45:9932–49. doi: 10.1109/TPAMI.2023.3237871

28. Zhang P, Yu X, Wang C, Zheng J, Ning X, Bai X. Towards effective person search with deep learning: a survey from systematic perspective. *Pattern Recognit*. (2024) 152:110434. doi: 10.1016/j.patcog.2024.110434

29. Dalakleidi KV, Papadelli M, Kapolos I, Papadimitriou K. Applying image-based food-recognition systems on dietary assessment: a systematic review. *Adv Nutr*. (2022) 13:2590–619. doi: 10.1093/advances/nmac078

30. Ismail N, Malik OA. Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. *Inf Process Agric*. (2022) 9:24–37. doi: 10.1016/j.inpa.2021.01.005

31. Dhanya V, Subeesh A, Kushwaha N, Vishwakarma DK, Kumar TN, Ritika G, et al. Deep learning based computer vision approaches for smart agricultural applications. *Artif Intell Agric*. (2022) 6:211–29. doi: 10.1016/j.aiia.2022.09.007

32. Mavani NR, Ali JM, Othman S, Hussain M, Hashim H, Rahman NA. Application of artificial intelligence in food industry-a guideline. *Food Eng Rev*. (2022) 14:134–75. doi: 10.1007/s12393-021-09290-z

33. Yin H, Yi W, Hu D. Computer vision and machine learning applied in the mushroom industry: a critical review. *Comput Electron Agric*. (2022) 198:107015. doi: 10.1016/j.compag.2022.107015

34. Ren B, Wang Z. Strategic priorities, tasks, and pathways for advancing new productivity in the Chinese-style modernization. *J Xi'an Univ Finance Econ*. (2024) 37:3–11.

35. Ma P, Li A, Yu N, Li Y, Bahadur R, Wang Q, et al. Application of machine learning for estimating label nutrients using USDA Global Branded Food Products Database, (BFPD). *J Food Compos Anal*. (2021) 100:103857. doi: 10.1016/j.jfca.2021.103857

36. Oliveira Chaves L, Gomes Domingos AL, Louzada Fernandes D, Ribeiro Cerqueira F, Siqueira-Batista R, Bressan J. Applicability of machine learning techniques in food intake assessment: a systematic review. *Crit Rev Food Sci Nutr*. (2023) 63:902–19. doi: 10.1080/10408398.2021.1956425

37. Prananto JA, Minasny B, Weaver T. Near infrared (NIR) spectroscopy as a rapid and cost-effective method for nutrient analysis of plant leaf tissues. *Adv Agron*. (2020) 164:1–49. doi: 10.1016/bs.agron.2020.06.001

38. Astill J, Dara RA, Campbell M, Farber JM, Fraser ED, Sharif S, et al. Transparency in food supply chains: a review of enabling technology solutions. *Trends Food Sci Technol*. (2019) 91:240–7. doi: 10.1016/j.tifs.2019.07.024

39. Singh V. Current challenges and future implications of exploiting the omics data into nutrigenetics and nutrigenomics for personalized diagnosis and nutrition-based care. *Nutrition*. (2023) 110:112002. doi: 10.1016/j.nut.2023.112002

40. Andraos S, Wake M, Saffery R, Burgner D, Kussmann M, O'Sullivan J. Perspective: advancing understanding of population nutrient-health relations via metabolomics and precision phenotypes. *Adv Nutr*. (2019) 10:944–52. doi: 10.1093/advances/nmz045

41. Yadav P, Menon N, Ravi V, Vishvanathan S, Pham TD. EfficientNet convolutional neural networks-based android malware detection. *Comput Secur*. (2022) 115:102622. doi: 10.1016/j.cose.2022.102622

42. He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans Geosci Remote Sens*. (2022) 60:1–15. doi: 10.1109/TGRS.2022.3144165

43. Shi C, Zhao S, Zhang K, Wang Y, Liang L. Face-based age estimation using improved Swin Transformer with attention-based convolution. *Front Neurosci*. (2023) 17:1136934. doi: 10.3389/fnins.2023.1136934

44. Zhu L, Lee F, Cai J, Yu H, Chen Q. An improved feature pyramid network for object detection. *Neurocomputing*. (2022) 483:127–39. doi: 10.1016/j.neucom.2022.02.016

45. Shao W, Min W, Hou S, Luo M, Li T, Zheng Y, et al. Vision-based food nutrition estimation via RGB-D fusion network. *Food Chem*. (2023) 424:136309. doi: 10.1016/j.foodchem.2023.136309

46. Ma P, Lau CP, Yu N, Li A, Sheng J. Application of deep learning for image-based Chinese market food nutrients estimation. *Food Chem*. (2022) 373:130994. doi: 10.1016/j.foodchem.2021.130994

47. Jiang L, Qiu B, Liu X, Huang C, Lin K. DeepFood: food image analysis and dietary assessment via deep model. *IEEE Access*. (2020) 8:47477–89. doi: 10.1109/ACCESS.2020.2973625

48. VijayaKumari G, Vutkur P, Vishwanath P. Food classification using transfer learning technique. *Global Transit Proc*. (2022) 3:225–9. doi: 10.1016/j.gltp.2022.03.027

49. Desai R, Gada A, Bari P. Image-based food calorie estimation using incremental learning. In: *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*. Bhubaneswar: IEEE (2022), p. 69–74. doi: 10.1109/MLCSS57186.2022.00021