Design of an Intelligent Dining Recommendation System for Nutritional Intake Based on Multimodal Recognition

Hao Chen
Nanjing University of Posts and Telecommunication College of Integrated Circuit Science and Engineering (College of Industry-Education Integration Nanjing), China

Abstract: Within the broader context of rising chronic diseases linked to dietary imbalances in China and the rapid advancement of artificial intelligence, this research focuses on developing an intelligent dining recommendation system based on multimodal nutrient intake recognition. Existing approaches are constrained by significant limitations, including the absence of robust multimodal nutritional quantification models, substantial spatio-temporal data fusion challenges, and considerable estimation errors caused by factors like ingredient occlusion, cooking transformations, and individual absorption variability. To address these deficiencies, this study employs a multimodal recognition framework integrating visual (RGB-D), depth, and skeleton data, combined with knowledge graph embedding and dynamic user interest modeling using attention mechanisms. Key findings demonstrate that the proposed system, through the construction of a Chinese Nutrient Map (CNM-NutriMap), achieves markedly lower error rates in calorie, protein, and fat estimation compared to traditional methods. Furthermore, it shows potential for enhancing dietary compliance in hypertension management and generating significant health economic savings. The manuscript's unique value lies in its theoretical contribution of a cross-domain "behavior-nutrition-physiology" correlation model and its practical innovation in providing a scalable, precision nutrition intervention tool for public health.

Keywords: Multimodal; Intelligent; Nutritional

1. Introduction

With the rapid advancement of artificial intelligence (AI), intelligent food recommendation systems have shown considerable potential for personalized dietary management. Multimodal information fusion, as a key enabling technology, has made significant progress in domains such as behavior recognition, object classification, and health intervention. Prior studies demonstrate that multimodal learning effectively overcomes the limitations of single-modal approaches. For example, the integration of RGB, skeleton, and optical flow data improves accuracy in human behavior recognition, while the combination of knowledge graphs with collaborative filtering algorithms enhances user preference modeling in dietary recommendation systems. These advances have accelerated the development of intelligent recommendation systems. Specifically, deep learning–based visual perception techniques have enabled the recognition of Chinese cuisine, whereas knowledge graph embedding methods address data sparsity and cold-start challenges, thereby supporting personalized dietary recommendations [1-2].

2. Research Background

From a national nutrition and health perspective, Chinese residents face pronounced dietary imbalances. Urban populations consume fat-derived energy well above recommended levels, the contribution of grain-based foods to overall energy intake has declined, and micronutrient deficiencies remain widespread. Such imbalances directly contribute to the increasing prevalence of chronic diseases, including overweight/obesity, hypertension, and diabetes. Among children and adolescents, selective eating behaviors contribute to high rates of abnormal body composition, while groups with higher socioeconomic status also exhibit elevated overweight prevalence. These trends highlight the urgent need for precision nutrition interventions [5-7].

Technologically, multimodal recognition provides novel approaches to addressing

nutritional monitoring challenges. Early studies relied primarily on single-modality data-such as RGB image-based dish recognition and collaborative filtering algorithms built on user rating matrices-but faced substantial limitations: visual recognition systems were sensitive to lighting conditions and occlusion, resulting in significant errors in nutritional estimation; meanwhile, traditional recommendation methods struggled with data sparsity, frequent cold-start problems, and static modeling constraints[2,4]. Recent advances in multimodal fusion have demonstrated promising solutions. For example, Jia Chen et al. proposed a width-learning approach that integrates RGB-D data to improve food recognition accuracy and reduce training time; Shao Bangli's team combined visual and audio modalities to enhance interactive command recognition; and Yin Yifan's research in tactile sensing enabled the quantitative characterization of food textures[23]. These technological developments support the design of intelligent, personalized catering management systems, although critical technical bottlenecks must still be overcome to achieve precise nutritional monitoring and dynamic dietary recommendations[1,8-9].

3. Research Significance

3.1 Theoretical Significance

This study develops a theoretical framework for multimodal nutritional recognition and dynamic recommendation, thereby extending the interdisciplinary boundaries between health informatics and artificial intelligence. Existing theoretical approaches exhibit three major limitations. First, multimodal fusion theories lack a unified paradigm, and the applicability of early-stage feature fusion versus late-stage decision fusion remains unclear. Second, nutritional quantification models rely primarily on empirical formulas and fail to establish explicit mappings between the three-dimensional morphology of ingredients and their nutritional components. Third, user interest modeling is still dominated by static collaborative filtering, which overlooks the temporal dynamics of health state evolution [3,10-11].

3.2 Practical Significance

Under the Healthy China strategy, the practical value of this system is reflected in three dimensions. First, in chronic disease prevention and control, real-time nutritional monitoring combined with DASH diet adaptation can reduce sodium intake among hypertensive patients, thereby lowering the incidence of cardiovascular events. The system's built-in dynamic prescription engine automatically optimizes recipes based on clinical indicators (e.g., blood pressure fluctuations), overcoming the static limitations of traditional interventions [12]. Second, in health management, the system utilizes the Chinese National Nutrition Database-developed from national nutrition survey data-to transform population-level nutrition strategies into personalized recommendations. By integrating smart refrigerator terminals with wearable devices, it improves household dietary compliance and mitigates selective eating behaviors in children [7,13-14]. Third, in food education, the system incorporates social cloud functionality to generate personalized educational plans based on users' dietary behavior profiles. This approach addresses the lack of a systematic food education mechanism in China and enhances residents' nutritional awareness [4,15].

4. Research Status of the Study Topic

Research on *Intelligent Food Recommendation Systems Based on Multimodal Recognition of Nutrient Intake* is currently developing along divergent trajectories worldwide. In China, studies emphasize the engineering implementation of multimodal perception technologies and have achieved notable progress in food recognition, behavioral analysis, and the integration of recommendation algorithms. Nevertheless, limitations persist in cross-modal generalization and dynamic adaptive modeling. Although specific international literature is not cited, global research trends demonstrate deeper engagement with frontier topics such as self-supervised cross-modal representation learning, explainable recommendation mechanisms, and edge computing deployment, with increasing attention to algorithmic transparency and privacy protection. Overall, the field is undergoing a paradigm shift from being "technology-driven" to "demand-driven," necessitating advances in both nutrition quantification accuracy and real-time personalized recommendation optimization [18].

4.1 Domestic Research Landscape

Domestic research primarily advances along

three directions: the optimization of multimodal perception technologies, the innovation of recommendation algorithms, and the integration of system applications. A comprehensive technological chain has been established, extending from data collection to personalized service delivery, thereby providing a robust foundation for the development of intelligent food recommendation systems [19].

4.1.1 Multimodal perception and nutritional recognition technologies

This research direction seeks to overcome technical bottlenecks in food and behavior recognition under complex scenarios by enhancing system robustness through the fusion of multi-source data, including visual, depth, and skeleton information. Jia Chen et al. proposed a multimodal information fusion framework based on width learning, employing RGB-D data augmentation to improve recognition of food variety, portion size, and morphological variations[20]. They further established cross-modal semantic associations between color and depth information using typical correlation analysis, thereby meeting real-time application requirements [1]. Mu Zhijia et al. emphasized that multimodal learning requires the integration of image, motion capture, and audio data to accurately identify food types and eating sequence features, thus providing foundational inputs for nutritional modeling [16]. Wang Cailing et al. experimentally validated the effectiveness of multimodal data (RGB, depth, skeleton, infrared) in capturing human eating behaviors, demonstrating stability under complex lighting and occlusion conditions [3]. Finally, Du Xiang et al. highlighted the limitations of single-image modalities and advocated for the fusion of sensory data, such as hyperspectral imaging, to enable joint analysis of food appearance and composition, thereby enhancing the distinguishability of similar ingredients [10].

4.1.2 Recommendation algorithms and knowledge modeling

Current research emphasizes knowledge-driven approaches and dynamic interest modeling to address data sparsity and cold-start challenges. Geng Huacong et al. constructed a dietary knowledge graph to represent complex relationships among recipes, ingredients, nutritional components, and constitutional types. By quantifying dish-level semantic similarity through embedded representation learning and incorporating it into a collaborative filtering framework, they significantly enhanced recommendation accuracy [2]. Zhang Xinyu critiqued traditional recommendation methods for overlooking evolving user interests and introduced an attention mechanism to capture temporal patterns in user behavior sequences for dynamic interest modeling [11]. Sheng Shiwang further integrated Traditional Chinese Medicine constitution classification theory with the Apriori algorithm, generating health-oriented menus based on constitution assessment and historical preference analysis [17].

4.2 Summary and Review of Research Status

4.2.1 Research gaps and limitations

Despite substantial progress in multimodal recognition and intelligent recommendation, critical gaps remain. Multimodal nutritional quantification models have yet to be established. Although advancements in food classification have been achieved through multimodal recognition-such as Jia Chen's width-learning approach and Wu Hang's visual perception system, which enhance food recognition accuracy-nutritional computation continues to rely on linear conversions from standard food composition tables [1,4]. Du Xiang observed that cooking-induced morphological changes in food introduce significant errors in traditional image analysis methods. Yin Yifan further demonstrated that a single visual modality is insufficient to characterize key food texture parameters, including viscosity and elasticity [9-10]. Currently, multidimensional perception systems integrating visual (color and texture), tactile (chewing resistance), and spectral (near-infrared composition detection) data are lacking, particularly with limited capability to accurately deconstruct the nutritional composition of complex Chinese dishes[21].

4.3 Research Challenges and Future Directions

4.3.1Core challenge analysis

Current research encounters technical bottlenecks primarily in two areas. First, multimodal data fusion challenges arise due to substantial dimensional differences among visual (high-dimensional images), tactile (time-series waveforms), and acoustic (spectral) data, spanning two to three orders of magnitude. Additionally, temporal delays between hand movements and food intake result in significant

synchronization errors in existing methods. Moreover, evaluation metrics for assessing the relative contribution of each modality to nutritional estimation are currently lacking [1,3,9]. Second, challenges in nutritional quantification modeling persist. Factors such as ingredient occlusion, changes in cooking form, seasoning penetration, and individual absorption variability significantly affect estimation accuracy. Existing approaches exhibit notable limitations: they rely exclusively on linear extrapolation of visible portions to address ingredient occlusion, fail to establish mapping models linking cooking form to nutrient components, cannot detect internal seasoning distribution, and neglect physiological parameters of digestion and metabolism, thereby limiting the capacity to account for individual absorption differences [9-10].

Table 1. Quantitative Analysis of Nutritional Components

Influencing Factors	Error Range	Limitations of Existing Solutions
Ingredient Occlusion	25%-40%	Linear estimation based solely on visible portions
Changes in cooking form	30%-50%	No established shape-to-ingredient mapping model
Seasoning penetration rate	45%-60%	Internal penetration distribution cannot be detected
Individual absorption variation	15%-25%	Digestive and metabolic physiological parameters ignored

4.3.2Future research directions

To address the challenges outlined above, potential breakthroughs may be pursued in the following directions. First, an adaptive multimodal fusion framework should be established, accompanied by a modality importance assessment module. The contribution of each modality can be quantified through typical correlation analysis, as demonstrated in Jia Chen's research[22]. A gated feature alignment mechanism can be designed to resolve spatio-temporal asynchrony, enabling the effective integration of multimodal data by weighting and fusing visual and acoustic features through learnable parameters [1]. Second, three-dimensional nutritional reconstruction techniques should be developed. By combining RGB-D cameras with near-infrared spectroscopy, as suggested in Du Xiang's research, food volume density models can be constructed. For example, integrating 3D point cloud segmentation with near-infrared spectroscopy improves the accuracy of solid food portion estimation. Fluid dynamics simulation is suitable for estimating fat distribution in soups and broths, while micro-CT can be employed to assess the porosity of baked goods [10]. V. Innovation Points and Research Value.

4.4 Theoretical Innovations

This study addresses existing theoretical limitations through three key innovations. First, it proposes a multimodal nutritional quantification model that employs cross-modal feature decoupling to decompose food characteristics into appearance invariance (illumination-invariant texture), structural invariance (rotation-invariant geometry), and composition-sensitive features (nutrient-associated vectors). Concurrently, the study constructs the first Chinese Nutrient Map (CNM-NutriMap), encompassing over 2,000 dishes across eight major Chinese culinary traditions. This mapping framework significantly outperforms traditional methods in annotating nutrients such as calories, protein, and fat [4,10], as shown in Table 2.

Table 2. Multimodal Nutrient Quantification Model

Nutrient	Allowed Error	Traditional Method Error
Calories	<5%	12%-25%
Protein	<8%	15%-30%
Fat	<10%	20%-45%

4.5 Research Value System

4.5.1Theoretical value

The theoretical contributions of this study are evident in three aspects. First, it advances multimodal learning theory by introducing a modality contribution quantification metric, which addresses the challenge of selecting optimal multimodal fusion strategies and provides a theoretical foundation for modal integration across diverse scenarios[1,3]. Second, it contributes to health informatics by establishing a cross-domain correlation model linking "behavior–nutrition–physiology," thereby bridging the theoretical gap between dietary recommendations and health interventions and offering new perspectives for interdisciplinary research at the intersection of health informatics and artificial intelligence [11-12].

4.5.2Application value

The applied value of this research is primarily reflected in two aspects. First, it supports chronic

disease management. In interventions for hypertension, the system enhances adherence to the DASH diet and achieves a significant reduction in mean systolic blood pressure, thereby providing an effective tool for dietary management of hypertensive patients [12]. Second, it generates health economic benefits. Estimates indicate that nationwide implementation could produce substantial cost savings for diabetes, cardiovascular disease, and obesity-related conditions, contributing to reductions in national healthcare expenditures [5-6], as shown in Table 3.

Table 3. Health Expenditure

Disease Type	Cost Savings*/ Person-Year	National Scale (¥billion/year)
Diabetes	¥2,800	420
Cardiovascular Disease	¥3,650	548
Obesity-Related	¥1,950	293

5. Literature Review Summary

Research in the field of intelligent catering over the past five years has followed a clear evolutionary trajectory. The focus has shifted from Wu and Hang's single-modal visual recognition system in 2019 to Jia Chen's multimodal fusion (width-learning) technology in 2022, and further to Zhang Xinyu's dynamic recommendation optimization (attention mechanism) in 2023. This trajectory reflects a progression from single-technology approaches to multi-technology integration and from static to dynamic methodologies [1,4,11].

Current research faces two principal gaps. First, multimodal nutritional decomposition models tailored for Chinese cuisine are lacking, particularly analytical capabilities for complex scenarios such as oil-salt permeation and ingredient co-variation. Second, existing recommendation systems struggle to achieve health-state-driven strategy adaptation. For example, when users transition from fitness phases (requiring high protein) to sugar-control phases (requiring low carbohydrates), system adjustments can lag by three to seven days [10-11]. These gaps constrain the practical applicability of intelligent dining systems in real-world settings, necessitating breakthroughs through cross-modal alignment algorithms and adaptive recommendation architectures.

The spatio-temporal fusion framework and dynamic perception engine proposed in this study are specifically designed to address these core gaps. They have the potential to advance health management from experience-based guidance toward precision intervention paradigms, thereby offering new pathways for the development of intelligent dietary recommendation systems.

References

[1] Jia Chen, Liu Huaping, Xu Xinyi, et al. Multimodal Information Fusion Based on Width Learning Method [J]. Journal of Intelligent Systems, 2019, 14(1): 150–157.

[2] Geng, H. C., Liang, H. T., & Liu, G. Z. (2021). Dietary recommendation algorithm based on knowledge graph and collaborative filtering. Computer & Modernization, (8): 24-29.

[3] Wang Cailing, Yan Jingjing, Zhang Zhidong. Research Review on Human Behavior Recognition Methods Based on Multimodal Data [J]. Computer Engineering and Applications, 2024, 60(9): 1-18

[4] Wu Hang. Research and Implementation of a Visual Perception-Based Dietary Health Management System [D]. Jinan: Shandong University, 2024

[5] Gu Jingfan. Interpretation of the "Report on the Status of Nutrition and Chronic Diseases among Chinese Residents (2015)" [J]. Chinese Journal of Nutrition, 2016, 38(6): 525-529

[6] Li L, Rao KQ, Kong LZ, et al. China National Nutrition and Health Survey 2002 [J]. Chinese Journal of Epidemiology, 2005, 26(7): 478-484

[7] Xu Yatao. Study on Physical Development Status and Influencing Factors Among Chinese Children and Adolescents [D]. Shanghai: East China Normal University, 2019

[8] Shao Bangli, Zhu Yin, Zhu Run, et al. A multimodal human-machine intelligent interaction method for smart home device control [J]. Journal of Forest Engineering, 2021, 6(4): 190-196

[9] Yin Yifan. Research on Robot Tactile Sensing and Multimodal Perception Technology [D]. Nanchang: Nanchang University, 2024

[10] Du Xiang, Qian Jiahe, Nie Wennan, et al. Research Progress of Modern Image Analysis Methods in Traditional Chinese Medicine Identification [J]. Chinese Journal of Pharmacy, 2025, 60(14): 1479-1485.

[11] Zhang Xinyu. Research and Development of a Knowledge Graph-Based Healthy Diet

Recommendation System [D]. Tianjin: Tianjin University of Science and Technology, 2023.
[12] Sun Ting. Research on Exercise and Dietary Interventions for Hypertension and Methods to Improve Compliance [D]. Hefei: University of Science and Technology of China, 2025.
[13] Gong Weiyan, Yuan Fan, Ding Caicui, et al. Design and Development of a National Nutrition and Health Assessment System [J]. Chinese Journal of Nutrition, 2025, 47(2): 108-112
[14] Li Nan, Jian Yuxuan, Bi Zusong, Gong Lei, Liu Zihao, Fan Jiale. Design of an Intelligent Refrigerator Recommendation System [J]. Internet of Things Technology, 2022, (12): 125-126
[15] Tang Hongtao, Liu Rui, Xia Rui, et al. Current Status of Food Education Practices Domestically and Internationally [J]. Chinese Food and Nutrition, 2020, 26(1): 5-8.
[16] Mou Zhijia, Fu Yaru. Review of Multimodal Learning Analysis Research [J]. Journal of Distance Education, 2021, 31(6): 23-31.
[17] Sheng Shiwang. Development of a Personalized Intelligent Diet Recommendation System [D]. Hangzhou: Zhejiang University of Science and Technology, 2015
[18] Cui Xiaohui, Li Wei, Gu Chengchun. Big Data and Artificial Intelligence Technologies in Food Science [J]. Chinese Journal of Food Science, 2021, 21(2): 1-8
[19] Liu Rui, Xu Luhui. Research on Intelligent Food Recommendation System Based on Flink [J]. Information Technology and Informatization, 2022, (10): 204-207
[20] Zhan Jianhao, Wu Hongwei, Zhou Chengzu, Chen Xiaoqiu, Li Xiaochao. A Review of Deep Learning-Based Multimodal Fusion Methods for Behavior Recognition [J]. Computer Systems Applications, 2023, 32(1): 41–49.
[21] Li Hongliang, Liu Yuliang, Liao Wenhui, Huang Mingxin, Zhang Shuo, Jin Lianwen. Optical Character Recognition in the Era of Large Models: Current Status and Outlook [J]. Chinese Journal of Image and Graphics, 2025, 30(6): 2023-2050
[22] Hu Yong. Research on Information Fusion Applications in Pattern Recognition [D]. Hefei: Hefei University of Technology, 2007.
[23] Liu Yongtao. Research on Human Behavior Recognition Based on Deep Learning [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022.

