Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 4 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i04.59905

PAPER

Multimodal Fusion of Image and Recipe Features for Nutrient Estimation of Complementary Foods

Nani Purwati[1,2] (✉), R. Rizal Isnanto[1], Martha Irene Kartasurya[1]

[1]Universitas Diponegoro, Semarang, Indonesia

[2]Universitas Bina Sarana Informatika, Yogyakarta, Indonesia

nani.npi@bsi.ac.id

ABSTRACT

The assessment of the nutritional content of complementary foods for infants aged 6–24 months is still largely done manually, which is time-consuming and prone to errors. This study proposes an automated nutritional content prediction model based on a multimodal approach that integrates food images and recipe texts. The experiment was conducted using the ComFoodID25 dataset, which consists of 2,783 images of complementary foods, complete with information on ingredients, processing methods, and ten types of nutrients. Visual features were extracted using pre-trained ResNet50, while text features were obtained using IndoBERT, and then both modalities were combined through a multilayer perceptron (MLP) architecture. The evaluation results showed that the multimodal model produced low error values for most nutrients, with an MAE value below 1 for the majority of nutrients and an overall PMAE value of 2.55%. Additionally, the high coefficient of determination values indicates a strong correlation between the predicted and reference values. These findings suggest that the proposed multimodal approach is effective and reliable for automatically estimating the nutritional content of complementary foods and has the potential to support artificial intelligence-based complementary food monitoring and recommendation systems.

KEYWORDS

multimodal learning, nutrition prediction, multilayer perceptron (MLP) fusion, food analysis, recipe features

1 INTRODUCTION

The appropriate provision of complementary foods is an important step in preventing malnutrition, which is one of the main risk factors for stunting [1] [2]. To support this practice, accurate methods of assessing nutritional content are needed to ensure the adequacy of the food provided. However, conventional methods for assessing the nutritional content of food typically involve manually weighing food ingredients and searching for nutritional information from various sources

Purwati, N., Isnanto, R. R., Kartasurya, M. I. (2026). Multimodal Fusion of Image and Recipe Features for Nutrient Estimation of Complementary Foods. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(4), pp. 140–154. https://doi.org/10.3991/ijoe.v22i04.59905

Article submitted 2025-12-03. Revision uploaded 2026-01-15. Final acceptance 2026-01-16.

separately [3]–[5]. This approach requires considerable time and resources, making it inefficient for large-scale or repeated assessments [3], [4].

Currently, image-based food recognition and nutritional assessment have made significant progress alongside developments in deep learning and computer vision techniques. This technology offers a promising solution for more accurate monitoring of nutritional intake and effective health management [6]–[14]. This system can estimate macronutrients and calories, providing users with valuable dietary insights [15]–[17]. Deep learning models, particularly convolutional neural networks (CNNs), have been widely used for food image classification and recognition, achieving high accuracy in identifying foods and estimating their nutritional content [14], [18].

Image-based nutritional assessment can improve the accuracy of nutritional assessment compared to traditional self-report methods, which are often inaccurate and time-consuming [4], [19], [20]. Image-based techniques enable automatic food recognition, portion size estimation, and nutritional analysis by leveraging deep learning algorithms and computer vision [13]. Although image-based nutritional assessment methods offer a more accurate and efficient solution than conventional methods, this approach still faces a number of technical and practical challenges that need to be overcome in order to be widely implemented. One of the main challenges is image quality and variability. Poor image quality, such as blurred images or inadequate lighting, can reduce the accuracy of food detection and nutritional content estimation [11], [12]. In addition, the variability in the shape, color, and appearance of food that arises during the preparation and consumption process adds to the complexity of the recognition and classification stage [11], [19].

Although various deep learning-based approaches have shown promising results in food classification and nutrient content estimation [21]–[24], most research still focuses on the recognition of common foods and does not fully accommodate the complexity of typical foods such as complementary foods, which have high visual similarity and diverse ingredient compositions. In addition, many existing models focus on classifying food based solely on visual features, which can lead to inaccuracies due to the high visual similarity between different food items [13], [20], [25]. Visual similarities between different food items and variations within a food class can hinder classification performance [13], [25]–[27]. Models that rely solely on visual features are often insufficiently accurate for proper nutritional assessment [19], [28].

Various datasets have been developed to support image-based food recognition and nutritional assessment research with diverse focuses and scopes. MedGRFood (42,880 images, 132 classes) highlights Mediterranean cuisine [29], CamerFood10 focuses on Sub-Saharan African cuisine [27], and MFOOD-32 includes 6,400 images of Moroccan cuisine [27]. VIPER-FoodNet (VFN) [13] and Food Portion Benchmark (FPB) [30] expand studies on Western foods and portion size estimation, whereas FoodNExTDB [8], NutriNet [31], and Taiwan Food Dataset [32] provide more extensive nutritional and visual data across cultures. Although diverse, most of these datasets still focus on common foods and do not yet cover the complex and locally unique characteristics of complementary foods in Indonesia.

This study proposes a multimodal model for predicting the nutritional content of complementary foods by combining image features from ResNet50 and text representations of recipes from IndoBERT. The two modalities are fused using a MLP fusion architecture to produce more accurate nutritional estimates than image-based approaches alone. This approach enables the model to capture the context of ingredients and cooking processes that cannot be seen from images, thereby significantly improving prediction quality.

Unlike most previous studies, which generally focused only on estimating macronutrients such as energy, carbohydrates, protein, and fat [33]–[37], this study also

expands the scope of assessment to include micronutrients that play an important role in the growth and development of children aged 6–24 months, including iron, zinc, calcium, vitamin A, vitamin C, and vitamin E.

Thus, this study makes three main contributions: (1) proposing a domain-specific multimodal framework tailored to Indonesian complementary feeding, incorporating local food characteristics and cooking practices through image and recipe information; (2) extending prior multimodal nutrient estimation studies—largely limited to macronutrient prediction—by enabling the estimation of both macro- and micronutrients using efficient MLP-based late fusion; and (3) establishing a practical foundation for culturally grounded AI-based nutritional monitoring systems for children aged 6–24 months.

2 RELATED WORK

Research in the field of image-based nutritional assessment has developed rapidly alongside advances in computer vision technology and deep learning. [38]. CNN, like GoogLeNet, ResNet, and EfficientNet, has been widely used for food image classification due to its high accuracy in recognizing various food categories [39]–[43]. Models such as Model-44 have demonstrated impressive training and validation accuracy, demonstrating their practical efficacy in real-world applications [14]. The creation of large, annotated food image datasets is essential for training deep learning models. Datasets such as Food-101, UEC-FOOD100, and ChinaFood-100 provide extensive data for various food categories, improving the model's ability to accurately recognize and estimate nutrients [14], [38], [44], [45]. However, most of these studies are still oriented towards general foods, while the context of complementary foods for breastfed infants (MP-ASI), with their complex variety of ingredients and textures, has not been explored extensively.

Over time, research has shifted towards multimodal integration to enrich food feature representation. Vision-based nutrition estimation has benefited from multimodal feature fusion, such as RGB-D fusion networks, which integrate visual and depth information to improve the accuracy of nutritional assessment [34] [37] [46]. The Visual-Ingredient Feature Fusion (VIF2) method combines visual and ingredient features to improve nutritional estimation, highlighting the importance of ingredient information in predicting nutritional value [1]. This approach has been proven to improve recognition performance by capturing the semantic context of food. However, its application is still limited to certain types of food and has not been designed for the nutritional assessment needs of children, especially for complementary foods that have high visual similarity between menus. Multimodal learning and feature fusion strategies have also been applied across diverse non-food domains, demonstrating the general applicability of multimodal frameworks beyond food-related tasks [47]–[49].

The integration of multimodal features in food research is a growing trend that improves the representation and understanding of food characteristics. Food detection, segmentation, and classification are highly challenging tasks due to the high visual variation and natural deformability of food objects [50]. Modern Vision-Language Models (VLMs) and deep learning approaches have shown promising potential but still face difficulties in recognizing visually similar food categories, particularly when it comes to distinguishing subtle differences in cooking methods or very slight variations in appearance [8]. The system needs to handle a variety of real-world conditions, such as overlapping objects, inconsistent lighting, and different food presentations [51].

Based on these research gaps, this study introduces a multimodal approach for automated nutritional assessment of complementary foods, where the primary novelty lies in the formulation of a domain-specific multimodal framework tailored to Indonesian complementary feeding. Distinct from prior studies that mainly focus on Western-style

foods or generic food datasets, this work integrates visual food appearance with recipe-level textual information derived from Indonesian local foods and cooking practices. Image features are extracted using a pre-trained ResNet50 model, while textual representations are obtained using IndoBERT, which is well-suited for Indonesian-language recipes. Although a standard MLP-based late fusion strategy is adopted, the main contribution of this study lies in the design of a multimodal nutrient estimation framework for local complementary foods, including the prediction of both macronutrients and critical micronutrients—such as iron, zinc, calcium, vitamin A, vitamin C, and vitamin E—that are essential for the growth and development of children aged 6–24 months.

3 METHOD

This study focuses on developing a multimodal approach-based nutritional prediction model, which combines visual information from complementary food images and textual information from recipes. The main stages of the research are summarized in the following flowchart (see Figure 1), which illustrates the process from data pre-processing, feature extraction, multimodal feature fusion, and model training using a feature fusion-based MLP architecture to the model performance evaluation stage.

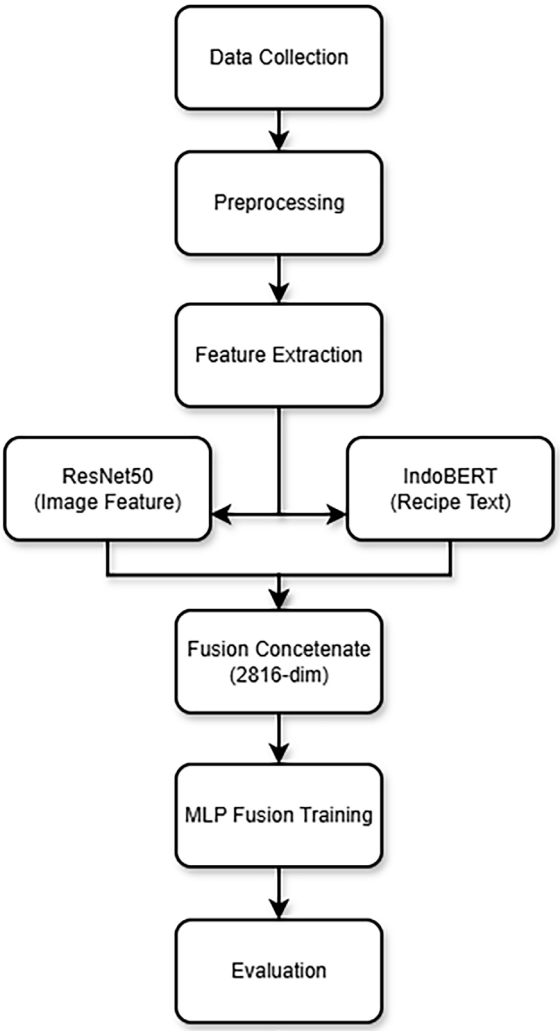

Fig. 1. Research framework

3.1 Data collection

This study uses ComFoodID25, a multimodal dataset designed to represent Indonesian complementary foods for children aged 6–24 months. Developed based on the 2023 Indonesian Ministry of Health complementary feeding guidelines [52] [53] and validated by nutrition experts using the NutriSurvey application, the dataset comprises 2,783 food images across 25 categories, as illustrated in Figure 2. Each sample includes a food image, recipe information (ingredients and cooking methods), and nutritional values for ten nutrients: energy, carbohydrates, fat, protein, calcium, vitamin A, vitamin C, vitamin E, zinc, and iron. Nutritional ground-truth values are computed using a standardized recipe-based calculation process, where ingredient-level nutritional values are aggregated according to portion size and national nutritional reference standards and subsequently reviewed by nutritionists to ensure consistency and validity.

Fig. 2. ComFoodID25 sample image display

3.2 Preprocessing

The pre-processing stage is carried out to prepare all data before entering feature extraction and model training. The image data were resized to 224×224 pixels and

normalized based on the ImageNet mean and standard deviation. Data augmentation techniques, including random resized crop, horizontal flip, and color jitter, were applied exclusively during the training phase to improve model robustness and generalization.

For text data, the ingredients and cooking instructions components are combined into a single *recipe_text and* then tokenized using the IndoBERT tokenizer so that it can be processed as language model input. For the nutrition labels, all values were normalized using StandardScaler, and specifically for vitamin A, a log1p transformation was applied to stabilize the distribution during training, followed by an expm1 operation at the evaluation stage. These steps ensure that the data are provided in an optimal and consistent format for the multimodal training process.

3.3 Feature extraction

At the feature extraction stage, images and text are processed separately using two pre-trained models. Complementary food images are extracted using ResNet50, which functions as a frozen feature extractor so that its weights are not retrained. The image is processed up to the final layer before being fully connected to produce a 2048-dimensional visual representation. Meanwhile, the recipe text is processed using IndoBERT, and the embedding corresponding to the CLS token is used as the textual representation, capturing the global contextual information of the entire recipe and producing a 768-dimensional feature vector. The CLS token is widely used for sentence-level and document-level representation in transformer-based models. Although alternative pooling strategies, such as mean pooling over token embeddings, were considered conceptually, the CLS token was selected due to its computational efficiency and suitability for global text representation in regression tasks. These two feature vectors are then stored in NumPy format as input for the next stage of model fusion and training.

3.4 Multilayer perceptron fusion training

The fused feature vectors are used as input to a MLP consisting of two hidden layers with 1024 and 256 units, respectively, employing ReLU activation and dropout, followed by an output layer with 10 units to predict all nutritional values. The model was trained exclusively on extracted multimodal features using the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 1×10^{-5}. Training was performed with a batch size of 64 for up to 100 epochs, with convergence monitored using the validation loss and early stopping applied when no improvement was observed for 12 consecutive epochs. A ReduceLROnPlateau scheduler (patience = 5, factor = 0.5) and Smooth L1 loss were used, and all experiments were conducted on an NVIDIA Tesla T4 GPU.

After prediction, all outputs were transformed back to their original scales. Specifically, inverse StandardScaler was applied to all nutrients, while vitamin A values were restored using the expm1 function to reverse the log1p transformation applied during preprocessing. This ensures that all predicted values are expressed in their original nutritional units, such as kcal, grams, milligrams, or IU, enabling direct comparison with the reference values.

3.5 Evaluation

The evaluation process was conducted to assess the accuracy of the model in predicting ten types of complementary food nutrients in the validation data. The evaluation used four regression metrics, namely mean absolute error (MAE) to measure the average absolute difference between the predicted value and the actual value; root mean squared error (RMSE) to assess quadratic errors that are more sensitive to large errors, percentage mean absolute error (PMAE); which is calculated as the ratio of MAE to the average actual nutrient value and interpreted in percentage form; and the coefficient of determination (R^2) to evaluate the overall goodness-of-fit between the predicted and ground-truth nutrient values. This combination of metrics provides a comprehensive overview of the model's performance in predicting nutritional values.

4 RESULT AND DISCUSSION

4.1 ComFoodID25 dataset

The ComFoodID25 dataset comprises 2,783 images of local Indonesian complementary foods, along with textual information on ingredients, cooking instructions, and nutritional values for ten nutrients. The dataset covers 25 food categories representing typical complementary foods for breastfed infants. For model development, the data are split into 80% training and 20% validation at the sample level, where each image–recipe pair is treated as an independent instance. Figure 3 illustrates the distribution of training and validation samples across the 25 complementary food categories, confirming the consistent application of the 80:20 split for each class.

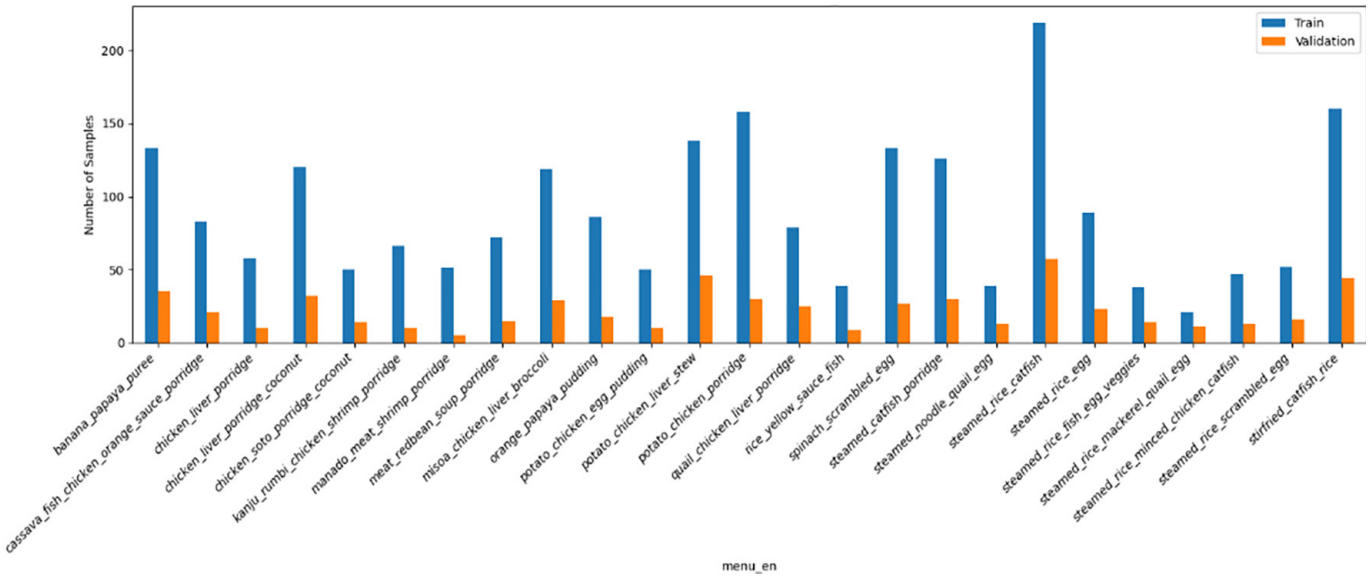

Fig. 3. Train vs. validation data distribution per class

4.2 Multimodal evaluation

A multimodal model was developed by combining visual features from ResNet50 and textual features from IndoBERT, which were then processed through the MLP Fusion architecture. This approach allows the model to utilize information from two modalities simultaneously so that nutritional content predictions do not only depend on the appearance of the food but also on recipe information such as ingredients, cooking methods, and dish composition.

Figure 4 shows that both training loss and validation loss decrease sharply in the first 10–15 epochs and then decline more gradually until they reach a plateau around epoch 40. The training and validation loss curves move very closely without any significant divergence, indicating that the model does not overfit and has good generalization. Validation performance improved consistently during training, marked by a decrease in MAE from 53.6 in the first epoch to around 2.6 at the end of training. Minor fluctuations after epoch 20 are a normal characteristic of multi-output regression with a highly variable range of nutritional values. Overall, the training pattern shows that the fusion model (image + text) successfully learned nutritional representations in a stable and efficient manner.

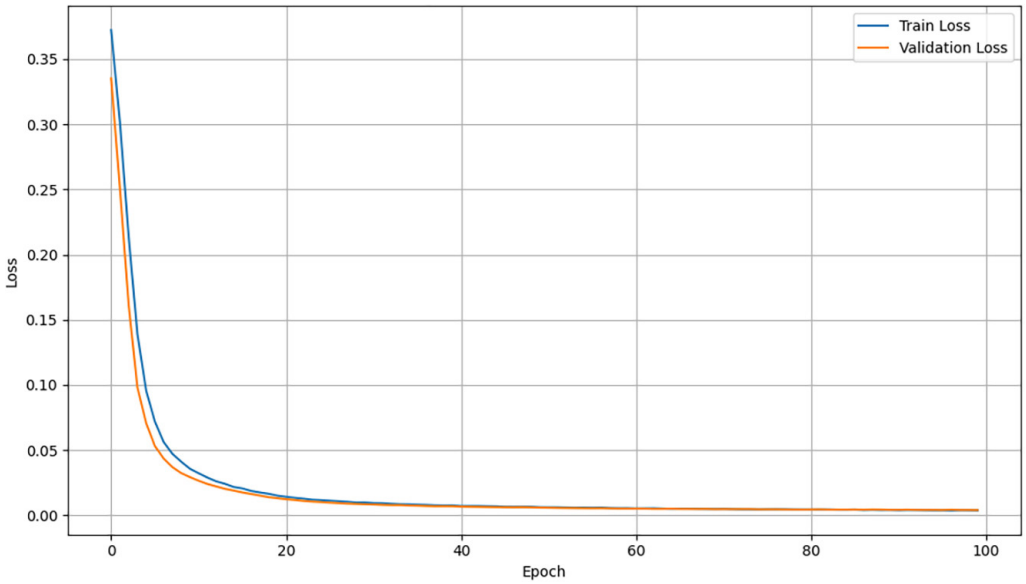

Fig. 4. Training and validation loss over epochs

The scatter plot visualization Figure 5 shows that the model is able to predict most nutrients with high precision, as indicated by the distribution of points that almost completely follow the diagonal line on each graph. The very small MAE values for energy, carbohydrates, fat, protein, calcium, vitamin E, vitamin C, zinc, and iron confirm this high accuracy, while the slight deviation seen in vitamin A is due to a much larger range of values compared to other nutrients, so that the absolute error appears higher even though the proportion remains small. Overall, the model shows consistent, robust, and reliable predictive performance.

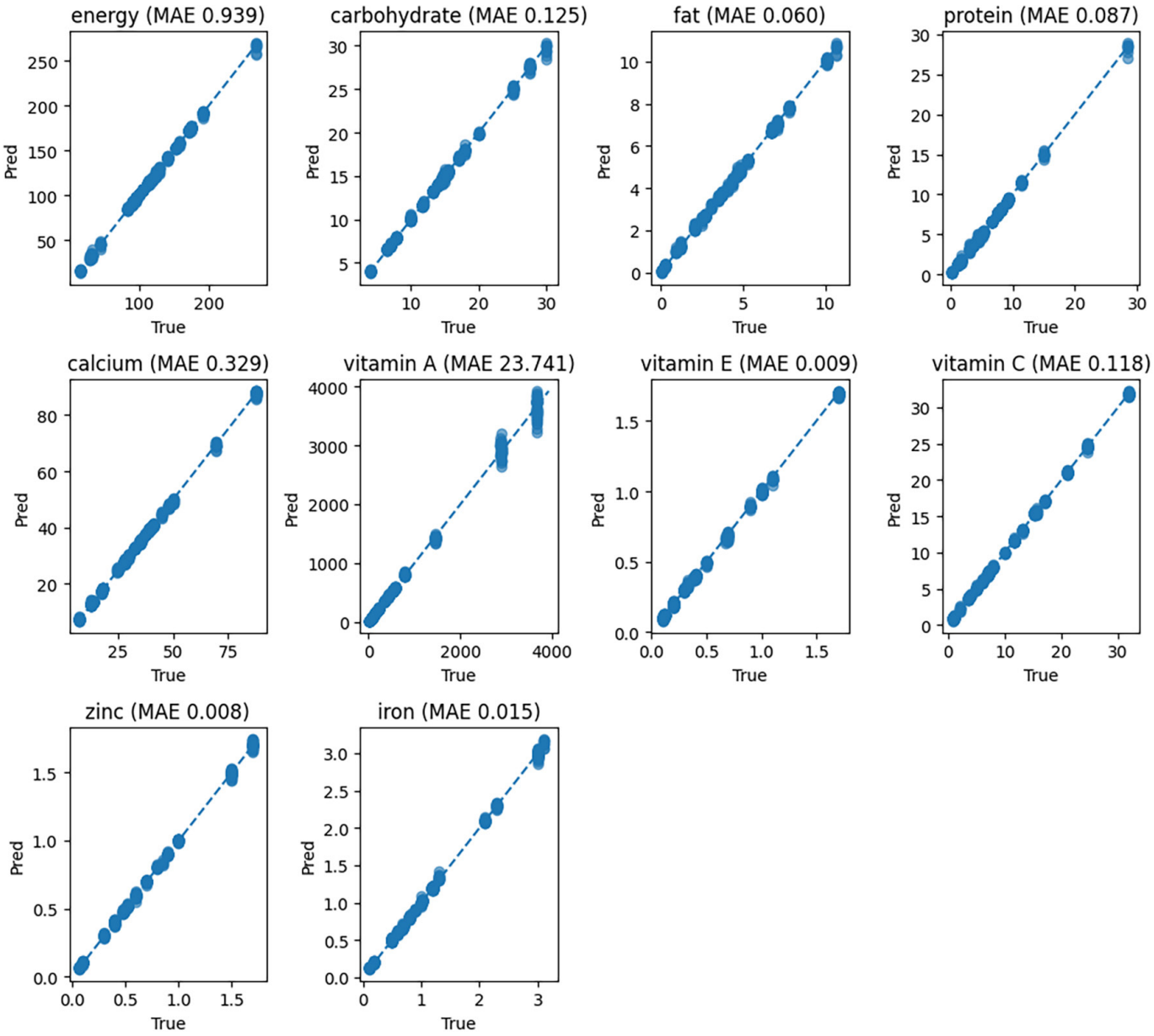

Fig. 5. Scatter plot visualization

Table 1 shows that vitamin A is the only nutrient with a noticeably higher error (MAE = 23.741; RMSE = 58.130), which is reasonable given its wide value range and its strong dependence on specific ingredients that are not always visually apparent or explicitly described in the recipe. In contrast, all other nutrients exhibit low and consistent error values, indicating stable prediction performance. In addition, the high overall coefficient of determination (R^2 = 0.999) demonstrates a strong goodness-of-fit between the predicted and reference nutrient values, further confirming the accuracy and robustness of the proposed multimodal model.

Table 1. Multimodal model evaluation matrix results table

Nutrition	MAE	RMSE	PMAE
Energy (kcal)	0.938	1.149	0.788
Carbohydrates (g)	0.122	0.183	0.859
Fat (g)	0.058	0.082	1.234
Protein (g)	0.090	0.153	1.317
Calcium (g)	0.314	0.433	0.960
Vitamin A (RE)	23.741	58.130	2.947
Vitamin C (mg)	0.008	0.011	1.604
Vitamin E (mcg)	0.110	0.148	0.971
Zinc (mg)	0.008	0.012	1.227
Iron (mg)	0.016	0.023	1.240
Overall MAE			**2.851**
Overall PMAE			**2.555%**
R^2			**0.999027**

4.3 Discussion

Previous studies on food nutrition estimation have primarily focused on visual or RGB-D information and have reported PMAE values in the range of 17–19%, reflecting the inherent challenges of estimating nutritional content from visual cues alone. For example, Shao et al. [34] achieved a PMAE of 18.5% with the RGB-D FusionNet model. Meanwhile, Zhao et al. [36] reported a PMAE of 17.06% using a single-image-based segmentation and regression approach. Furthermore, Nian et al. [37] integrated material information into an RGB-D-based framework (IMIR-Net), achieving a PMAE of approximately 17.4%.

Although direct quantitative comparison across these studies is not strictly equivalent due to differences in datasets, nutrient scopes, and evaluation settings, the results obtained in this study indicate that integrating visual features with recipe-based textual information can substantially reduce prediction errors on the evaluated complementary food dataset. Specifically, the proposed multimodal model achieves an overall PMAE of 2.55% on the ComFoodID25 dataset, highlighting the potential advantages of leveraging structured recipe information for nutritional estimation tasks. To provide a transparent overview, Table 2 summarizes the characteristics and reported performance of representative related studies alongside the proposed approach.

Table 2. Comparison analysis with the previous study

Study	Year	Dataset	Method	Predicted Nutrients	Results
Shao et al. [34]	2023	Nutrition5k (RGB-D)	RGB-D FusionNet	Energy, Mass, Fat, Carbohydrates, Protein	PMAE 18.5%
Zhao et al. [36]	2024	Nutrition5k	Food segmentation and RGB-based regression	Energy, Carbohydrates, Fat, Protein	PMAE 17.06%

(Continued)

Table 2. Comparison analysis with the previous study *(Continued)*

Study	Year	Dataset	Method	Predicted Nutrients	Results
Nian et al. [37]	2024	Nutrition5k	IMIR-Net	Energy, Mass, Fat, Carbohydrates, Protein	PMAE 17.4%
Baseline		ComFoodID25 (Image only)	ResNet50	10 macro and micronutrients	**Overall PMAE** 53.29%
Multimodal		ComFoodID25 local food dataset and recipe text	MLP Fusion Architecture	10 macro and micronutrients	**Overall PMAE 2.55%**

5 CONCLUSION

This study proposes a multimodal framework for predicting the nutritional content of complementary foods by integrating visual information and recipe-based textual data. The model combines features extracted using ResNet50 and IndoBERT through an MLP-based fusion architecture.

Experimental results show that the multimodal approach consistently outperforms the image-only baseline across all evaluated nutrients, substantially reducing both absolute and relative prediction errors. The integration of textual information provides important contextual cues related to ingredients and preparation methods, leading to more accurate nutritional estimation. Although vitamin A remains more challenging to predict due to its uneven distribution and ingredient dependency, the overall model demonstrates stable performance and strong agreement between predicted and reference values.

Despite the use of a single training–validation split, the findings confirm that multimodal learning offers a promising and effective direction for estimating the nutritional content of complementary foods for infants aged 6–24 months. Future work will explore more comprehensive evaluation strategies to further assess model robustness and generalization.

6 REFERENCES

[1] H. Qi, B. Zhu, C.-W. Ngo, J. Chen, and E.-P. Lim, "Advancing food nutrition estimation via visual-ingredient feature fusion," in *ICMR 2025 – Proceedings of the 2025 International Conference on Multimedia Retrieval*, 2025, pp. 1091–1099. https://doi.org/10.1145/3731715.3733269

[2] Z. Shao, G. Vinod, J. He, and F. Zhu, "An end-to-end food portion estimation framework based on shape reconstruction from monocular image," in *Proceedings – IEEE International Conference on Multimedia and Expo*, vol. 2023, 2023, pp. 942–947. https://doi.org/10.1109/ICME55011.2023.00166

[3] G. Vinod, Z. Shao, and F. Zhu, "Image based food energy estimation with depth domain adaptation," in *Proceedings – 5th International Conference on Multimedia Information Processing and Retrieval, MIPR 2022*, 2022, pp. 262–267. https://doi.org/10.1109/MIPR54900.2022.00054

[4] S.-T. Cheng, Y.-J. Lyu, and C. Teng, "Image-based nutritional advisory system: Employing multimodal deep learning for food classification and nutritional analysis," *Appl. Sci.*, vol. 15, no. 9, p. 4911, 2025. https://doi.org/10.3390/app15094911

[5] Z. Shao *et al.*, "An integrated system for mobile image-based dietary assessment," *AI & Food'21: Proceedings of the 3rd Workshop on AIxFood*, 2021, pp. 19–23. https://doi.org/10.1145/3475725.3483625

[6] K. Moumane, I. El Asri, T. Cheniguer, and S. Elbiki, "Food recognition and nutrition estimation using MobileNetV2 CNN architecture and transfer learning," in *Proceedings – SITA 2023: 2023 14th International Conference on Intelligent Systems: Theories and Applications*, 2023. https://doi.org/10.1109/SITA60746.2023.10373725

[7] L. Jiang, B. Qiu, X. Liu, C. Huang, and K. Lin, "DeepFood: Food image analysis and dietary assessment via deep model," *IEEE Access*, vol. 8, pp. 47477–47489, 2020. https://doi.org/10.1109/ACCESS.2020.2973625

[8] S. Romero-Tapiador *et al.*, "Are vision-language models ready for dietary assessment? Exploring the next frontier in AI-Powered food image recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2025, pp. 430–439. https://doi.org/10.1109/CVPRW67362.2025.00047

[9] R. Krutik, C. Thacker, and R. Adhvaryu, "Advancements in food recognition: A comprehensive review of deep learning-based automated food item identification," in *2024 2nd International Conference on Electrical Engineering and Automatic Control, ICEEAC 2024*, 2024. https://doi.org/10.1109/ICEEAC61226.2024.10576416

[10] E. J. Delp, Y. Han, J. He, M. Gupta, E. J. Delp, and F. Zhu, "Diffusion model with clustering-based conditioning for food image generation," in *MADiMa 2023 – Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management, Co-located with: MM 2023*, 2023, pp. 61–69. https://doi.org/10.1145/3607828.3617796

[11] B. Maharana, M. K. Goyal, and A. I. Abidi, "Bridging the gap: From food recognition to accurate weight estimation," in *ICDT 2025 – 3rd International Conference on Disruptive Technologies*, 2025, pp. 1157–1162. https://doi.org/10.1109/ICDT63985.2025.10986752

[12] Y. Han, S. K. Yarlagadda, T. Ghosh, F. Zhu, E. Sazonov, and E. J. Delp, "Improving food detection for images from a wearable egocentric camera," in *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 2021, 2021, no. 8. https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-286

[13] R. Mao, J. He, L. Lin, Z. Shao, H. A. Eicher-Miller, and F. Zhu, "Improving dietary assessment via integrated hierarchy food classification," in *IEEE 23rd International Workshop on Multimedia Signal Processing, MMSP 2021*, 2021. https://doi.org/10.1109/MMSP53017.2021.9733586

[14] S. Khawate, S. Gaikwad, Y. Davda, R. Shirbhate, P. Gham, and V. Borate, "Dietary monitoring with deep learning and computer vision," in *2025 International Conference on Computing Technologies and Data Communication, ICCTDC 2025*, 2025. https://doi.org/10.1109/ICCTDC64446.2025.11158839

[15] B. Kalivaraprasad, M. V. D. Prasad, and N. K. Gattim, "Deep learning-based food calorie estimation method in dietary assessment: An advanced approach using convolutional neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 1044–1050, 2024. https://doi.org/10.14569/IJACSA.2024.01503104

[16] V. Van Wymelbeke-Delannoy *et al.*, "A cross-sectional reproducibility study of a standard camera sensor using artificial intelligence to assess food items: The foodintech project," *Nutrients*, vol. 14, no. 1, p. 221, 2022. https://doi.org/10.3390/nu14010221

[17] D. Ganpisetty, C. R. Reddy, N. Ganpisetty, and J. Anitha, "Real-time food detection and nutritional tracking application for personalized health management using MobileNetV2," in *8th IEEE International Conference on Computational System and Information Technology for Sustainable Solutions, CSITSS 2024*, 2024. https://doi.org/10.1109/CSITSS64042.2024.10816813

[18] A. Peng, J. He, and F. Zhu, "Self-supervised visual representation learning on food images," in *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 35, 2023. https://doi.org/10.2352/EI.2023.35.7.IMAGE-269

[19] S. Zhang, V. Callaghan, and Y. Che, "Image-based methods for dietary assessment: A survey," *J. Food Meas. Charact.*, vol. 18, no. 1, pp. 727–743, 2024. https://doi.org/10.1007/s11694-023-02247-2

[20] S. Madhumitha, M. Magimaa, M. Maniratnam, and N. Neelima, "Dietary assessment and nutritional analysis using deep learning," *Lect. Notes Electr. Eng.*, vol. 844, pp. 11–21, 2022. https://doi.org/10.1007/978-981-16-8862-1_2

[21] N. Purandhar, S. Poojitha, S. M. Hussain, M. P. Chowdary, and M. Rafi, "Food recognition and calorie estimation in mixed food items using MobileNet," in *Proceedings of 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2025*, 2025, pp. 1014–1019. https://doi.org/10.1109/ICICV64824.2025.11085688

[22] C. Kiourt, G. Pavlidis, and S. Markantonatou, "Deep learning approaches in food recognition," in *Learning and Analytics in Intelligent Systems*, vol. 18, Greece: Springer Nature, 2020, pp. 83–108. https://doi.org/10.1007/978-3-030-49724-8_4

[23] B. Shah and H. Bhavsar, "Depth-restricted convolutional neural network—a model for Gujarati food image classification," *Vis. Comput.*, vol. 40, no. 3, pp. 1931–1946, 2024. https://doi.org/10.1007/s00371-023-02893-z

[24] A. Reethika, T. Jagadesh, and M. S. Kanivarshini, "Nutrition food recognition using deep learning algorithm for physically challenged human being," in *Deep Learning for Cognitive Computing Systems: Technological Advancements and Applications*, Department of Electronics and Communication Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India: De Gruyter, 2022, pp. 113–128. https://doi.org/10.1515/9783110750584-007

[25] X. Pan, J. He, and F. Zhu, "Muti-stage hierarchical food classification," in *MADiMa 2023 – Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management, Co-located with: MM 2023*, 2023, pp. 79–87. https://doi.org/10.1145/3607828.3617798

[26] X. Pan, J. He, and F. Zhu, "FMiFood: Multi-modal contrastive learning for food image classification," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing, MMSP 2024*, 2024. https://doi.org/10.1109/MMSP61759.2024.10743395

[27] T. R. Baban A Erep and L. Chaari, "mid-DeepLabv3+: A novel approach for image semantic segmentation applied to African food dietary assessments," *Sensors*, vol. 24, no. 1, p. 209, 2024. https://doi.org/10.3390/s24010209

[28] C.-F. Chung *et al.*, "Opportunities to design better computer vison-assisted food diaries to support individuals and experts in dietary assessment: An observation and interview study with nutrition experts," *PLOS Digit. Heal.*, vol. 3, no. 11, p. e0000665, 2024. https://doi.org/10.1371/journal.pdig.0000665

[29] F. S. Konstantakopoulos *et al.*, "GlucoseML mobile application for automated dietary assessment of mediterranean food," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2022, vol. 2022, pp. 1432–1435. https://doi.org/10.1109/EMBC48229.2022.9871732

[30] A. Sanatbyek *et al.*, "A multitask deep learning model for food scene recognition and portion estimation—the Food Portion Benchmark (FPB) dataset," *IEEE Access*, vol. 13, pp. 152033–152045, 2025. https://doi.org/10.1109/ACCESS.2025.3603287

[31] S. Mezgec and B. K. Seljak, "Nutrinet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, p. 657, 2017. https://doi.org/10.3390/nu9070657

[32] G. G. C. Lee *et al.*, "Single food image database: A comprehensive high quality image dataset for food recognition in artificial intelligence," in *IEEE International Conference on Electro Information Technology*, 2025, pp. 383–388. https://doi.org/10.1109/eIT64391.2025.11103677

[33] H. J. Wen, S. L. Wang, M. C. Li, and Y. L. Guo, "Aspergillus sensitization associated with current asthma in children in the United States: An analysis of data from the 2005–2006 NHANES," *Epidemiol. Health*, vol. 44, p. e2022099, 2022. https://doi.org/10.4178/epih.e2022099

[34] W. Shao *et al.*, "Vision-based food nutrition estimation via RGB-D fusion network," *Food Chem.*, vol. 424, no. February, p. 136309, 2023. https://doi.org/10.1016/j.foodchem.2023.136309

[35] K. Lee, "Multispectral food classification and caloric estimation using convolutional neural networks," *Foods*, vol. 12, no. 17, p. 3212, 2023. https://doi.org/10.3390/foods12173212

[36] Y. Zhao, P. Zhu, Y. Jiang, and K. Xia, "Visual nutrition analysis: Leveraging segmentation and regression for food nutrient estimation," *Front. Nutr.*, vol. 11, pp. 1–15, 2024. https://doi.org/10.3389/fnut.2024.1469878

[37] F. Nian, Y. Hu, Y. Gu, Z. Wu, S. Yang, and J. Shu, "Ingredient-guided multi-modal interaction and refinement network for RGB-D food nutrition assessment," *Digit. Signal Process. A Rev. J.*, vol. 153, p. 104664, 2024. https://doi.org/10.1016/j.dsp.2024.104664

[38] P. Ma *et al.*, "Image-based nutrient estimation for Chinese dishes using deep learning," *Food Res. Int.*, vol. 147, p. 110437, 2021. https://doi.org/10.1016/j.foodres.2021.110437

[39] S. Salma, M. Habib, A. Tannouche, and Y. Ounejjar, "Comparative analysis of convolutional neural network architectures for poultry meat classification," *IAES Int. J. Artif. Intell.*, vol. 14, no. 5, pp. 3715–3723, 2025. https://doi.org/10.11591/ijai.v14.i5.pp3715-3723

[40] M. Sumanth, A. H. Reddy, D. Abhishek, S. V. Balaji, K. Amarendra, and P. V. V. S. Srinivas, "Deep learning based automated food image classification," in *Proceedings – 2024 2nd International Conference on Inventive Computing and Informatics, ICICI 2024*, 2024, pp. 103–107. https://doi.org/10.1109/ICICI62254.2024.00026

[41] J. Sultana, B. M. Ahmed, M. M. Masud, A. K. O. Huq, M. E. Ali, and M. Naznin, "A study on food value estimation from images: Taxonomies, datasets, and techniques," *IEEE Access*, vol. 11, pp. 45910–45935, 2023. https://doi.org/10.1109/ACCESS.2023.3274475

[42] D. Al-Rubaye and S. Ayvaz, "Deep transfer learning and data augmentation for food image classification," in *2022 Iraqi International Conference on Communication and Information Technologies, IICCIT 2022*, 2022, pp. 125–130. https://doi.org/10.1109/IICCIT55816.2022.10010432

[43] A. Singla, L. Yuan, and T. Ebrahimi, "Food/non-food image classification and food categorization using pre-trained GoogLeNet model," in *MADiMa 2016 – Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Co-located with ACM Multimedia 2016*, 2016, pp. 3–11. https://doi.org/10.1145/2986035.2986039

[44] K. V. Dalakleidi, M. Papadelli, I. Kapolos, and K. Papadimitriou, "Applying image-based food-recognition systems on dietary assessment: A systematic review," *Adv. Nutr.*, vol. 13, no. 6, pp. 2590–2619, 2022. https://doi.org/10.1093/advances/nmac078

[45] E. Tasci, "Voting combinations-based ensemble of fine-tuned convolutional neural networks for food image recognition," *Multimed. Tools Appl.*, vol. 79, nos. 41–42, pp. 30397–30418, 2020. https://doi.org/10.1007/s11042-020-09486-1

[46] Y. Han, Q. Cheng, W. Wu, and Z. Huang, "DPF-Nutrition: Food nutrition estimation via depth prediction and fusion," *Foods*, vol. 12, no. 23, p. 4293, 2023. https://doi.org/10.3390/foods12234293

[47] K. Yanai, T. Maruyama, and Y. Kawano, "A cooking recipe recommendation system with visual recognition of food ingredients," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 8, no. 2, pp. 28–34, 2014. https://doi.org/10.3991/ijim.v8i2.3623

[48] M. A. H. Saedan, M. Kassim, and A. F. Abd Aziz, "Biological butterfly characterization with mobile system using convolutional neural network (CNN) classify image," *Int. J. Interact. Mob. Technol.*, vol. 18, no. 7, pp. 125–138, 2024. https://doi.org/10.3991/ijim.v18i07.46267

[49] Q. A. Memon, "Multi-layered multimodal biometric authentication for smartphone devices," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 14, no. 15, pp. 222–230, 2020. https://doi.org/10.3991/ijim.v14i15.15825

[50] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato, "Retrieval and classification of food images," *Comput. Biol. Med.*, vol. 77, pp. 23–39, 2016. https://doi.org/10.1016/j.compbiomed.2016.07.006

[51] P. Panindre, P. K. Thummalapalli, T. Mandal, and S. Kumar, "Deep learning framework for food item recognition and nutrition assessment," in *6th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2025 – Proceedings*, 2025, pp. 1648–1653. https://doi.org/10.1109/ICMCSI64620.2025.10883519

[52] M. I. Kartasurya *et al.*, *Makanan Pendamping Asi (MP-ASI) Bagi Baduta*. Semarang: Undip Press, 2020.

[53] Kemenkes RI, "Makanan Lokal," *Buku Resep Makanan Lokal Bayi, Balita dan Ibu Hamil*, pp. 1–52, 2023.

7 AUTHORS

Nani Purwati received a master's degree in Computer Science from STMIK Nusa Mandiri. She is currently pursuing a doctoral degree in Information Systems at Diponegoro University. Her research interests include information systems, computer vision and image processing, biomedical image analysis, and user interface and user experience (UI/UX) design, particularly the application of artificial intelligence in healthcare systems and applications (E-mail: nani.npi@bsi.ac.id).

R. Rizal Isnanto obtained his bachelor's and master's degrees in electrical engineering at Gadjah Mada University, Yogyakarta, Indonesia. The doctoral degree was received at the Department of Electrical Engineering and Information Technology, Gadjah Mada University, in 2013. He received an honorary degree as a professor on June 1, 2023, as a professor in the field of image processing. He has held several positions, including Head of the Computer Engineering Department, Faculty of Engineering, Diponegoro University (2016–2021). He currently serves as Head of the Information Systems Doctoral Study Program at the Postgraduate School of Diponegoro University (2025–present). The study conducted to date is related to the fields of information systems, biomedical and biometric image processing, and pattern recognition (E-mail: rizal@ce.undip.ac.id).

Martha Irene Kartasurya is a medical doctor who received a master's degree in nutrition from Cornell University, USA, and a PhD in nutrition from the University of Queensland, Australia. She is currently serving as a Professor in the public health nutrition at the Public Health Nutrition Department, Faculty of Public Health, Diponegoro University, Semarang, Indonesia. Her research interests include maternal and child nutrition and maternal and child health (E-mail: marthakartasurya@live.undip.ac.id).