**JOIN (Jurnal Online Informatika)**
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 11 Number 1 | June 2026: 60-69
DOI: 10.15575/join.v11i1.1725

# Early Fusion of Visual and Ingredient Representations for Multimodal Food Classification

**Navira Rahma Salsabila[1], Adela Regita Azzahra[2], Fitri Utaminingrum[3], Barlian Henryranu Prasetio[4]**
[1,2,3,4]Faculty of Computer Science, Brawijaya University, Malang, Indonesia

**Article Info**

***Article history:***

Received June 29, 2025
Revised July 13, 2025
Accepted August 11, 2025
Published April 25, 2026

***Keywords:***

Artificial Intelligence
CLIP
Early Fusion
Food Recognition
Multimodal Classification

**ABSTRACT**

Identifying the most appropriate food dish based on available kitchen ingredients remains a practical yet challenging task in everyday life. To address this, this study specifically aims to develop an intelligent food classification system using a multimodal approach. We propose a multimodal food classification method that performs early fusion by combining visual and textual features extracted using the Contrastive Language–Image Pretraining (CLIP) model. Features from food images and ingredient lists are fused and classified through a two-layer multilayer perceptron. The model is evaluated on the Recipes5k dataset with 4,826 samples across 101 food categories. Results show that the proposed multimodal model achieves 91.32% accuracy, outperforming text-only (85.65%) and image-only (57.26%) baselines. The main contribution of this work lies in demonstrating the effectiveness of early fusion for combining cross-modal representations in food classification. Unlike prior methods, our model supports flexible inference with either text or image input, enabling practical real-world applications. These findings highlight the potential of multimodal learning for food recommendation systems, offering both accuracy and contextual relevance beyond unimodal approaches.

***Corresponding Author:***

**Fitri Utaminingrum**,
Faculty of Computer Science, Brawijaya University, Malang, Indonesia
Jl. Veteran No.10-11, Ketawanggede, Malang, Indonesia 65145
Email: f3_ningrum@ub.ac.id

## 1. INTRODUCTION

The task of selecting an appropriate dish based on available kitchen ingredients poses a recurring challenge in everyday life. This difficulty is further amplified by the limitations of manual recipe searches, which are time-consuming and often fail to provide contextually relevant results. With the advancement of Artificial Intelligence (AI), there has been a growing interest in developing intelligent and context-aware food recommendation systems. For instance, a model based on ResNet-V2 has demonstrated success in recognizing ingredients and suggesting suitable recipes without manual intervention [1]. Recent developments also include multimodal systems that combine refrigerator images with large language models (LLMs) to generate personalized recipes based on dietary preferences, time constraints, and portion sizes [2], [3]. These approaches offer a more adaptive and user-centered experience.

In this context, AI-based multimodal approaches have emerged as a promising solution, as they are capable of processing two primary data types: text (ingredient lists) and images (food visuals). Integrating both modalities enables systems to deliver more accurate and contextually relevant recommendations. However, building a robust food classification system remains challenging due to

high visual similarity between dishes, lighting variability, and cultural differences in food presentation. Unimodal systems those relying solely on text or images often fail to fully capture these complexities. For example, tomato soup and red bean soup may appear visually identical despite differing in composition and flavor [4], [5].

To overcome these limitations, multimodal learning has become a central focus in AI research. One of the most prominent models in this domain is Contrastive Language Image Pretraining (CLIP), developed by OpenAI [6]. CLIP maps visual and textual inputs into a shared semantic space, enabling more efficient cross-modal matching [7], [8]. It also supports zero-shot learning, handles noisy inputs, and performs well in multilabel classification tasks [9]. Introduced by Radford et al. [6], and trained on hundreds of millions of image–text pairs, CLIP has become foundational in vision–language research. Prior studies show that combining visual and textual features improves classification, especially for visually similar but semantically distinct foods [10].

CLIP has been benchmarked against other vision–language models. For example, Jia et al. (2021) proposed ALIGN; however, CLIP remains more widely adopted due to its open-source nature and strong generalization [11]. Applied Sciences (2025) found standard CLIP to be a strong baseline in food classification and nutritional estimation compared to a domain-specific variant, NutritionCLIP [12].

In the food domain, Kim et al. (2024) combined CLIP embeddings with large language models using transformer-based cross-attention, achieving strong results but with high complexity. In contrast, our approach uses lightweight early fusion by concatenating CLIP image and ingredient embeddings, followed by a simple classifier avoiding caption generation and complex attention making it more scalable for food recommendation systems.

Multimodal learning has also been applied in medical detection; for example, Setiawan et al. [13] used structured metadata for stunting prediction. While both involve multimodal integration, their approach uses tabular features, whereas ours focuses on semantic fusion of image and text using pretrained CLIP embeddings, within the domain of food classification and with an emphasis on efficient implementation.

This study proposes a multimodal food classification system using CLIP to extract image and ingredient embeddings, combined through lightweight early fusion via simple concatenation avoiding complex architectures. The fused features are classified with a two-layer neural network and evaluated on the Recipes5k dataset (101 categories) against unimodal baselines.

## 2. METHOD

This study proposes a practical multimodal food classification system using early fusion of image and ingredient embeddings via CLIP in five phases: (1) preprocessing: cleaning and resizing images, normalizing ingredient text; (2) feature extraction: encoding images and texts into 1024-dimensional vectors using CLIP; (3) embedding fusion: concatenating both vectors into a 2048-dimensional representation; (4) classification: applying a two-layer MLP (512-unit hidden layer, softmax output for 101 classes); and (5) evaluation: assessing accuracy, precision, recall, and F1-score on the validation set.

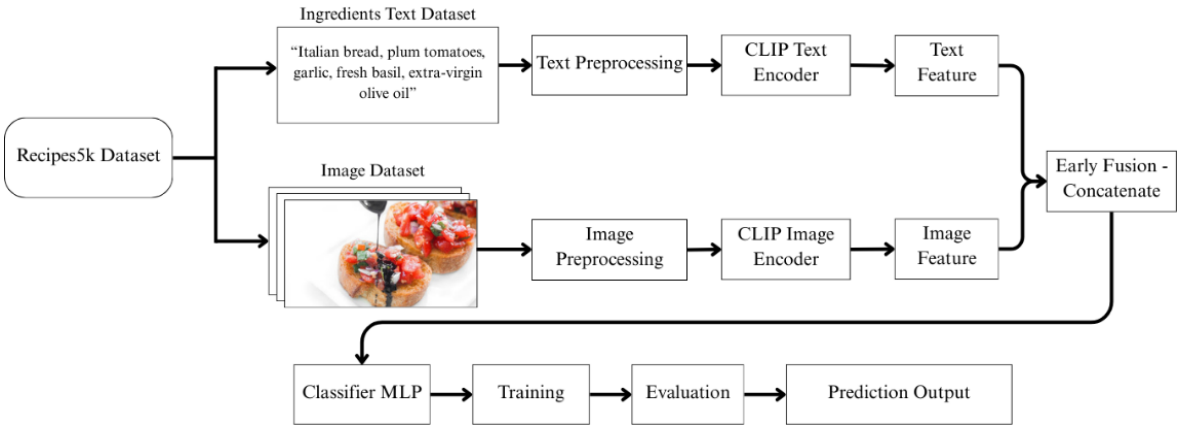


Figure 1. Research flow – Process from dataset preparation to CLIP feature extraction, early fusion, classification, and evaluation.

CLIP embeddings are precomputed to reduce training time. Early fusion is chosen for simplicity and computational efficiency, while CLIP provides robust, generalizable semantic representations. This method is fully empirical, integrating multimodal inputs through real experiments rather than theoretical derivation.

### *2.1. Dataset*

This study utilizes the Recipes5k dataset introduced by Bolaños et al., which was originally designed for multi-label ingredient recognition [14]. The dataset comprises recipes, each consisting of a single food image and a corresponding ingredient list, covering a variety of food categories from Food-101. Unlike its original purpose, which focused on detailed ingredient prediction, this study employs Recipes5k to develop a multimodal food classification system. In this framework, food images and ingredient lists are treated as two separate modalities processed by the model to enable two types of predictions: predicting the dish solely from the ingredient list, and identifying the dish from an image while simultaneously predicting its ingredients. The dataset is partitioned into 3,409 training samples, 634 validation samples, and 783 test samples, following the predefined split scenario provided by the dataset creators to ensure consistent usage across studies. All metadata, including image filenames, class labels, ingredient lists, and data split indicators, is provided in an annotated CSV file.

### *2.2. Preprocessing*

Preprocessing was performed separately for image and text data. Food images were resized to 224×224 pixels, normalized with the CLIP RN50 distribution [15], converted to tensors, and batched using PyTorch's DataLoader. Ingredient lists were tokenized with the CLIP tokenizer into fixed-dimensional embeddings aligned to the visual feature space [16]. As CLIP employs absolute positional encoding with a fixed input length, inputs were limited to 77 tokens, truncating longer sequences [17]. Batch tokenization excluded invalid or empty entries to maintain data integrity.

No additional image augmentation techniques such as random cropping, flipping, or color jitter were applied. This decision was based on two considerations. First, the CLIP model used for feature extraction was pretrained on unaugmented natural images, and applying augmentations could introduce a distribution shift. Second, the Recipes5k dataset already contains high intra-class visual variance, so preserving natural visual semantics ensures consistent embedding extraction. To reduce computational overhead, all image and text embeddings were precomputed using CLIP and saved as .pt files in PyTorch tensor format along with their class labels. This precomputation eliminates the need for repeated encoding during each training epoch, thereby accelerating the training process.

### *2.3. Feature Extraction with CLIP*

Feature extraction for both image and text modalities was performed using the pretrained CLIP model by OpenAI, which includes two frozen encoders: a ResNet-50 for images and a Transformer for text. These encoders project inputs into a shared 1024-dimensional semantic space, enabling effective cross-modal alignment for multimodal classification [18]. Operating in parallel, they allow joint representation of both modalities. Image preprocessing followed CLIP's standard resizing to 224×224 pixels and normalization using the RN50 distribution to maintain compatibility and input quality, which is essential for accurate classification [19]. As feature extraction is central to multimodal systems [20], early fusion was chosen due to its effectiveness in downstream CLIP applications [21].

As shown in Figure 2, each food image is encoded into a visual embedding, while the tokenized ingredient list is converted into a textual embedding. Leveraging CLIP's pretrained alignment, both embeddings are fused through concatenation into a unified representation. To accelerate training and reduce computational cost, all embeddings were precomputed and stored alongside their class labels.

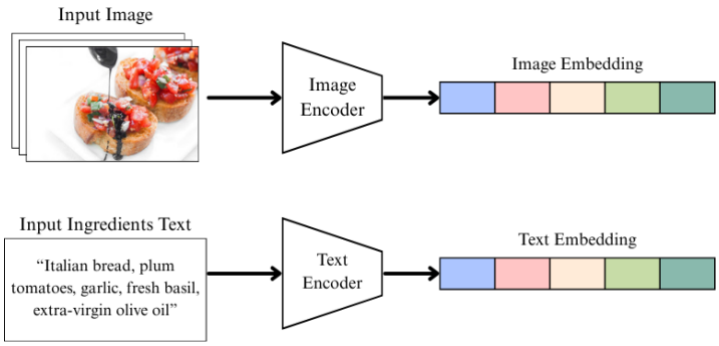


*Figure 2. CLIP-based early fusion – Image/text 1024-D, fused 2048-D, two-layer MLP.*

### *2.4. Multimodal Fusion Strategy*

This study adopts early fusion, combining visual and textual representations before classification. Food images are encoded with CLIP's ResNet-50 visual encoder, and ingredient texts with CLIP's Transformer text encoder. The resulting embeddings are concatenated into a fixed-length multimodal vector, with visual features capturing essential cues such as color, texture, and shape. [22]. Although no manual visual feature extraction is performed, the CLIP-generated embeddings inherently encapsulate rich semantic information from these attributes.

Early fusion is selected due to its implementation efficiency and its capacity to enable the model to directly learn cross-modal semantic relationships, provided that embeddings are appropriately normalized through pretraining [23]. This architecture also supports a more lightweight and computationally efficient model, resulting in improved accuracy and faster training times [12], [24].
The system is designed to remain flexible under real-world conditions where only one modality may be available. In such cases, the missing modality is substituted with a zero vector of equivalent dimensionality, preserving input structure without requiring architectural changes or retraining.

### *2.5. Model Architecture and Training Procedure*

The proposed classification model utilizes a fused representation of food images and ingredient texts. Each modality is independently encoded using CLIP encoders, with ResNet50 for image data and a Transformer for text data, producing 1024-dimensional embeddings. These embeddings are concatenated into a 2048-dimensional vector and passed through a two-layer MLP. The first layer reduces the dimensionality to 512 and applies ReLU. The second layer outputs logits corresponding to 101 food categories in the Recipes5k dataset.

This compact architecture was chosen for its computational efficiency and suitability for precomputed multimodal embeddings. To minimize training overhead, all CLIP-based image and text embeddings were computed in advance. The model was trained using a supervised learning setup with standard optimization and loss functions, as summarized in Table 1.

*Table 1. Model training parameters.*

| Parameter | Value |
|---|---|
| Loss Function | CrossEntropyLoss |
| Optimizer | Adam |
| Learning Rate | 1e-4 |
| Number of Epochs | 15 |
| Batch Size | 32 |

### *2.6. Evaluation Metrics*

To assess the performance of the developed classification model, four primary evaluation metrics are employed: accuracy, precision, recall, and F1-score. These metrics are used to evaluate how accurately and consistently the model performs in the context of multiclass classification [25]. The evaluation formulas used in this study are presented sequentially in Equations (1) through (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

p-ISSN: 2528-1682
e-ISSN: 2527-9165

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Equations (1)–(4) define the evaluation metrics used in this study: Accuracy is the proportion of correct predictions over total samples; Precision is the proportion of correct positive predictions over all predicted positives; Recall is the proportion of correct positive predictions over all actual positives; and F1-score is the harmonic mean of Precision and Recall, balancing false positives and false negatives. Evaluation is conducted both at the per-class level and across all classes using the macro average and weighted average approaches. The macro average computes the unweighted mean of the metric across all classes, treating each class equally regardless of sample size. In contrast, the weighted average assigns weights to each class based on the proportion of samples it contains [26]. This strategy provides a balanced assessment of model performance across both majority and minority classes.

### *2.7. Inference Design and Flexibility*

The classification model in this study is a multimodal system that combines information from food images and ingredient lists. Although trained using both modalities together, the model is designed to flexibly handle input in either visual or textual form, enabling practical use in diverse real-world scenarios. When the input is a food image, the model predicts the class using CLIP-based visual embeddings and presents the associated ingredient list. When the input is a textual ingredient list, the model generates a text embedding, predicts the food class, and shows three representative images from the predicted category. This dual-modality output enriches the classification experience with both contextual and semantic information.

This design illustrates that multimodal capabilities are not limited to the training phase, but are also applied during inference and result presentation. By delivering outputs in both visual and textual formats, the system not only enhances interpretability but also promotes user trust and transparency in the predictions.

## 3. RESULT AND DISCUSSION

This section presents a comprehensive evaluation of the proposed multimodal food classification system, which integrates food images and ingredient lists using an early fusion strategy. The main contribution of this study is the development of a lightweight and practical framework that leverages CLIP-based visual and textual embeddings, combined through simple concatenation and classified using a compact two-layer neural network. By using precomputed embeddings, the system enables efficient training while maintaining high performance. The following analysis presents the model's classification performance, compares it with unimodal baselines, and discusses its flexibility in handling input from either image or text, which reflects the robustness and applicability of the early fusion approach in practical food recognition scenarios.

### *3.1. Evaluation Results and Analysis of the Early Fusion Model*

The model was trained using an early fusion strategy that combines CLIP-generated embeddings from food images and ingredient texts. These 1024-dimensional embeddings were concatenated into a 2048-dimensional vector and passed through a two-layer MLP to classify 101 food categories from the Recipes5k dataset. Evaluated on the test set, the model achieved 91.32% accuracy, demonstrating strong classification performance even among visually similar food types. Full evaluation metrics are presented in Table 2.

The training and validation curves in Figure 3 show steady convergence for both loss and accuracy, with improvements gradually plateauing after the early epochs. As no significant gain was

observed in the final few epochs, the model was considered to have effectively converged. Therefore, training was stopped at 15 epochs to ensure optimal performance without unnecessary computation.

Table 2. Evaluation results of the early fusion model.

| Metric | Value (%) |
|---|---|
| Accuracy | 91.32 |
| Precision | 91.31 |
| Recall | 91.32 |
| F1-score | 90.72 |

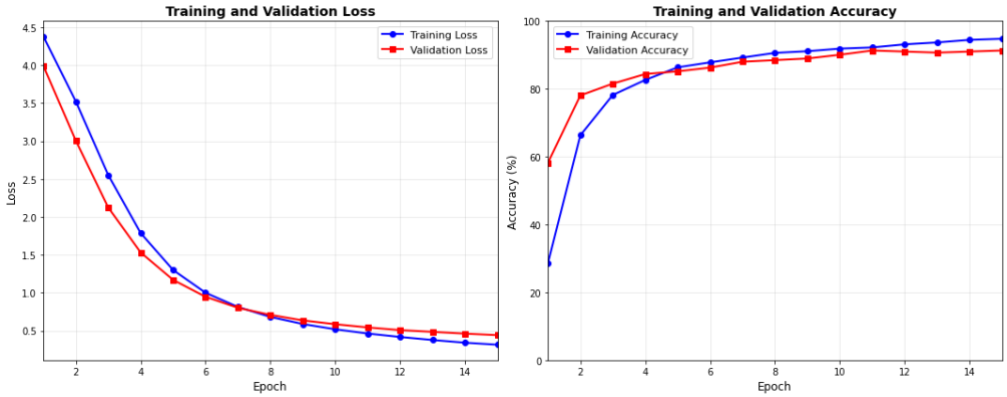


Figure 3. Training and Validation Loss/Accuracy – Model shows stable convergence with high accuracy and no overfitting.

To ensure interpretability and maintain visual clarity, Figure 4 presents the confusion matrix for the top 20 food classes with the highest sample frequencies in the test set. These classes were selected not only due to their prevalence in the dataset but also to offer a representative subset for evaluating model performance on frequently occurring and visually similar categories. Including all 101 classes in a single matrix would introduce excessive visual clutter and hinder meaningful analysis. By focusing on these 20 dominant categories, the confusion matrix more effectively highlights classification patterns and model behavior on critical food types.

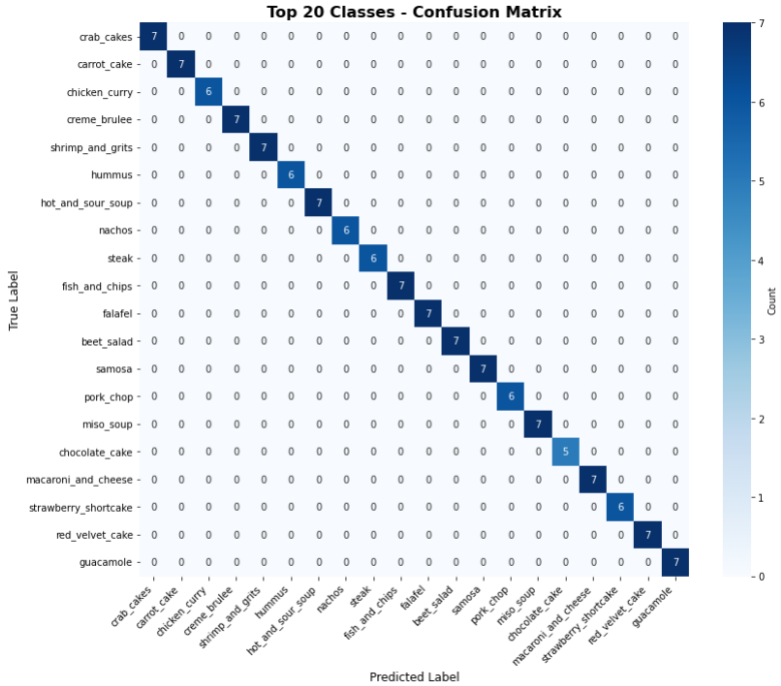


Figure 4. Confusion matrix of top 20 food classes shows clear diagonal, indicating high accuracy and class separability.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

The strong diagonal structure observed in the matrix not only confirms high classification accuracy but also highlights the model's capacity to form deeply aligned semantic representations. By jointly processing visual cues and ingredient semantics, the model moves beyond surface-level feature recognition and captures contextual meaning, which is essential for fine-grained classification tasks. These findings demonstrate the superiority of multimodal learning over unimodal approaches and establish the CLIP-based early fusion framework as a highly viable solution for real-world food recognition applications. The results also underline the potential of this method for broader implementation in intelligent culinary systems, multimodal recipe retrieval, and visual ingredient identification tools.

### *3.2. Comparison Between Multimodal and Unimodal Food Classification Approaches*

To assess the effectiveness of the multimodal approach, we compared it against two unimodal baselines: an image-only model and a text-only model based on ingredient descriptions. This comparison aimed to quantify the contribution of each modality and evaluate the benefit of integrating both. As shown in Table 3, the image-only model achieved 57.26% accuracy and 52.92% F1-score, indicating limited discriminative power when food items share similar visual traits. The text-only model performed considerably better, with 85.65% accuracy and 84.51% F1-score, leveraging the semantic richness of ingredients. The early fusion model outperformed both, achieving 91.32% accuracy and 90.72% F1-score.

Table 3. Performance comparison between multimodal and unimodal models.

| Approach | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Image-Only | 57.26% | 54.82% | 57.26% | 52.92% |
| Text-Only | 85.65% | 86.14% | 85.65% | 84.51% |
| **Early Fusion** | **91.32%** | **91.31%** | **91.32%** | **90.72%** |

These findings highlight that multimodal fusion leverages complementary strengths. Ingredient descriptions provide contextual information when visual cues are ambiguous, while images reinforce predictions when text is vague or incomplete. By modeling both modalities jointly, the early fusion model improves robustness and generalization, making it suitable for practical applications such as menu recommendation and content-based food retrieval.

### *3.3. Multimodal Inference Evaluation*

To evaluate the system's reasoning capabilities under real-world conditions, we conducted inference-based tests using both textual and visual inputs. In the text-only evaluation, the model was given either a single ingredient (e.g., "cheese") or a combination of ingredients (e.g., "chocolate", "vanilla", "sugar"). In response, it predicted plausible dishes such as cheese_plate, macaroni_and_cheese, and chocolate_ice_cream, demonstrating an understanding of semantic relationships between ingredients and food categories. These results, illustrated in Figure 5, confirm the model's ability to infer dish types based on varying levels of textual input complexity.

In a separate image-only evaluation, the system correctly identified a food photograph as bruschetta and inferred a relevant ingredient set including tomatoes, olive oil, garlic, and basil. This showcases the model's ability to perform visual recognition and semantic inference without text. As shown in Figure 6, the model maintains high predictive relevance across modalities. These findings highlight the system's cross-modal generalization and its potential for practical use in tasks such as recipe retrieval, menu recommendation, and intelligent dietary assistance. By leveraging shared embeddings, the model bridges visual and textual semantics, making it a robust foundation for adaptive AI-driven food applications.

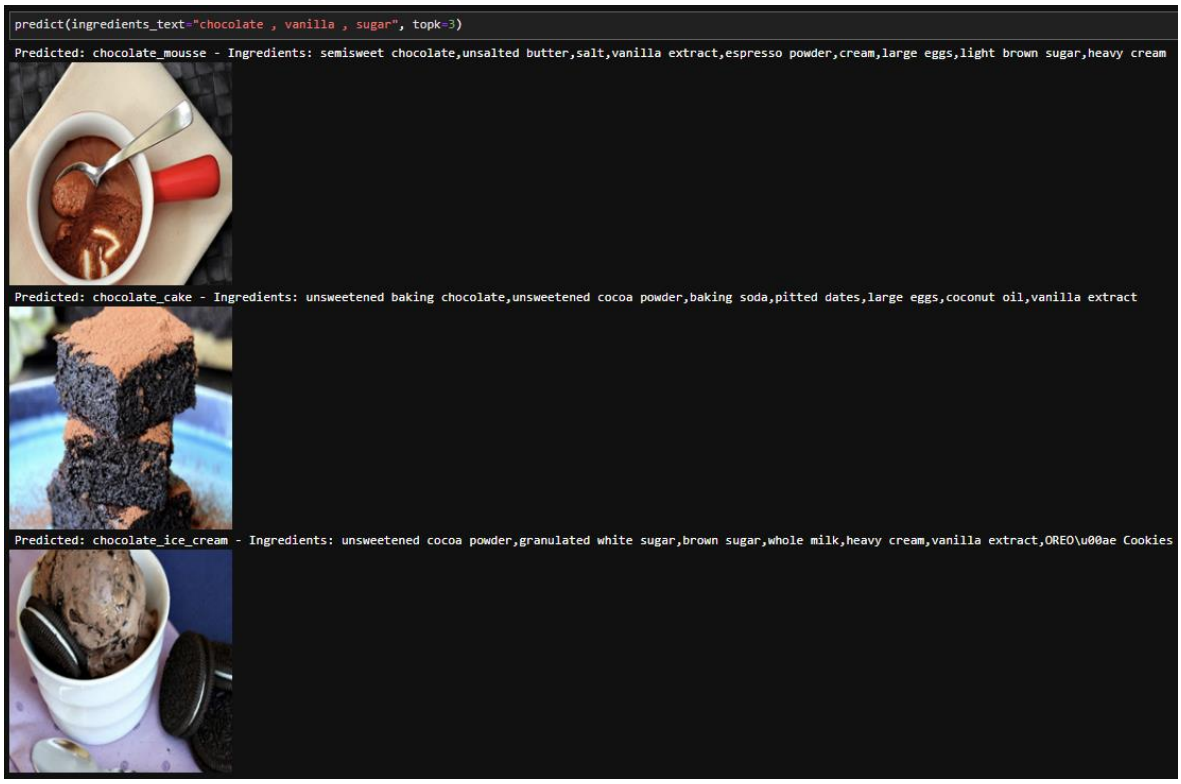


Figure 5. Prediction from ingredient inputs – Generates accurate dish predictions from ingredients only, showing strong text-based understanding.


Figure 6. Prediction from image input – Identifies dishes and predicts ingredients from images, showing cross-modal capability.

### *3.4. Discussion*

The experimental results demonstrate that combining image and ingredient features using an early fusion strategy significantly improves classification performance. Unlike unimodal models, the proposed multimodal approach effectively captures both visual appearance and semantic content, reducing confusion among visually similar dishes. The model also exhibits stable learning behavior, with consistent convergence in training and no signs of overfitting, which suggests strong generalization to unseen data. The confusion matrix supports this by showing a strong diagonal pattern, indicating the model's ability to distinguish between frequent and visually similar categories through effective semantic alignment.

In addition, the system shows flexibility during inference by producing reliable predictions even when only one modality, either image or ingredients, is available. This capability is particularly useful for real-world applications such as food tracking, recipe retrieval, or personalized menu recommendations, where complete multimodal input may not always be present. Overall, the combination of robust accuracy, generalization capability, and input flexibility underscores the effectiveness of the CLIP-based early fusion approach for practical multimodal food classification.

## 4. CONCLUSION

Food classification presents significant challenges due to overlapping visual characteristics and ambiguous ingredient descriptions, often leading to inaccurate predictions in unimodal models. To directly address this problem, this study evaluated a CLIP-based early fusion of visual and textual features. The proposed approach achieved 91.32% accuracy, outperforming image-only (57.26%) and text-only (85.65%) baselines, confirming that semantic-level fusion improves accuracy and context understanding. The system also performed well in real-world inference, accurately handling both ingredient-based and image-based queries, demonstrating practical applicability.

The primary contribution is a lightweight and scalable fusion method that leverages pretrained CLIP embeddings from both modalities without requiring attention mechanisms, handcrafted features, or complex architectures, making it efficient and deployable. Future research may explore adding modalities such as user preferences, dietary constraints, or cooking instructions, and adapting the model for real-time deployment in mobile or embedded systems. This study not only provides an effective solution to the research problem but also lays the groundwork for practical and extensible multimodal systems in culinary AI and related domains.

## ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Department of Computer Science, Universitas Brawijaya, for the academic and technical support provided throughout the research. Appreciation is also extended to the academic supervisor for the valuable guidance and constructive feedback during the system development. Access to the Recipes5k dataset and the publicly available CLIP model from OpenAI played a crucial role in supporting the experiments and evaluation of the proposed system.

## REFERENCES

[1] M. Ashraf *et al.*, "Improved Ingredients-based Recipe Recommendation Software using Machine Learning," in *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, Nov. 2023, pp. 509–514. doi: 10.1109/ICICIS58388.2023.10391164.

[2] D. Noever and S. E. M. Noever, "The Multimodal And Modular Ai Chef: Complex Recipe Generation From Imagery," *arXiv Prepr. arXiv2304.02016*, 2023, doi: 10.48550/arXiv.2304.02016.

[3] A. Saklani, S. Tiwari, and H. S. Pannu, "Deep attentive multimodal learning for food information enhancement via early-stage heterogeneous fusion," *Vis. Comput.*, vol. 41, no. 4, pp. 2461–2476, Mar. 2025, doi: 10.1007/s00371-024-03546-5.

[4] I. Gallo, G. Ria, N. Landro, and R. La Grassa, "Image and Text fusion for UPMC Food-101 using BERT and CNNs," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, Nov. 2020, pp. 1–6. doi: 10.1109/IVCNZ51579.2020.9290622.

[5] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo, and K.-M. Chao, "Text–image multimodal fusion model for enhanced fake news detection," *Sci. Prog.*, vol. 107, no. 4, Oct. 2024, doi: 10.1177/00368504241292685.

[6] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of Machine Learning Research*, 2021, vol. 139, pp. 8748–8763.

[7] B. Arpit, K. Kumar, and S. Singla, "Multimodal Deep Learning: Integrating Text and Image Embeddings with Attention Mechanism," in *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, May 2024, pp. 1–6. doi: 10.1109/AIIoT58432.2024.10574665.

[8] J. Guo, Y. Li, G. Cheng, and W. Li, "Based-CLIP early fusion transformer for image caption," *Signal, Image Video Process.*, vol. 19, no. 2, p. 112, Feb. 2025, doi: 10.1007/s11760-024-03721-0.

[9] F. Wang *et al.*, "MuMIC - Multimodal Embedding for Multi-Label Image Classification with Tempered Sigmoid," *Proc. 37th AAAI Conf. Artif. Intell. AAAI 2023*, vol. 37, pp. 15603–15611, 2023, doi: 10.1609/aaai.v37i13.26850.

[10] J.-H. Kim, N.-H. Kim, D. Jo, and C. S. Won, "Multimodal Food Image Classification with Large Language Models," *Electronics*, vol. 13, no. 22, p. 4552, Nov. 2024, doi: 10.3390/electronics13224552.

[11] C. Jia *et al.*, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 4904–4916, 2021.

[12] 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, Apr. 2025, doi: 10.3390/app15094911.

[13] Y. Setiawan, M. H. Z. Al Faroby, M. N. P. Ma'ady, I. M. W. A. Sanjaya, and C. V. C. Ramadhani, "Modality-based Modeling with Data Balancing and Dimensionality Reduction for Early Stunting Detection," *J. Online Inform.*, vol. 10, pp. 53–65, Apr. 2025, doi: 10.15575/join.v10i1.1495.

[14] R. Ismail and Z. Yuan, "Food ingredients recognition through multi-label learning," *Embed. Artif. Intell. Devices, Embed.*

*Syst. Ind. Appl.*, pp. 130–141, 2022, doi: 10.1201/9781003394440-10.

[15] G. Kwon and J. C. Ye, "CLIPstyler: Image Style Transfer with a Single Text Condition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022, vol. 2022-June, pp. 18041–18050. doi: 10.1109/CVPR52688.2022.01753.

[16] E. Kim, K. Shim, S. Chang, and S. Yoon, "Semantic Token Reweighting for Interpretable and Controllable Text Embeddings in CLIP," *EMNLP 2024 - 2024 Conf. Empir. Methods Nat. Lang. Process. Find. EMNLP 2024*, pp. 14330–14345, 2024, doi: 10.48550/arXiv.2410.08469.

[17] I. Najdenkoska, M. M. Derakhshani, Y. M. Asano, N. van Noord, M. Worring, and C. G. M. Snoek, "TULIP: Token-length Upgraded CLIP," pp. 1–24, 2025, [Online]. Available: http://arxiv.org/abs/2410.10034

[18] M. Mohammadi, M. Eftekhari, and A. Hassani, "Image-Text Connection: Exploring the Expansion of the Diversity Within Joint Feature Space Similarity Scores," *IEEE Access*, vol. 11, pp. 123209–123222, 2023, doi: 10.1109/ACCESS.2023.3327339.

[19] A. F. Rahmaniati and F. Utaminingrum, "Deep Learning Based Smart Wheelchair Navigation Optimization for Multi-Lighting Conditions," in *2024 4th International Conference on Robotics, Automation, and Artificial Intelligence, RAAI 2024*, 2024, pp. 295–300. doi: 10.1109/RAAI64504.2024.10949523.

[20] B. H. Prasetio, E. R. Widasari, and F. A. Bachtiar, "A Study of Machine Learning Based Stressed Speech Recognition System," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 4, pp. 31–42, 2022, doi: 10.22266/ijies2022.0831.04.

[21] A. Munthuli *et al.*, "Redefining the Classification of Extravasation Severity Using CLIP Linear Probe with Few-shot Instances," 2024. doi: 10.1109/EMBC53108.2024.10782522.

[22] A. Septiarini, A. Sunyoto, H. Hamdani, A. A. Kasim, F. Utaminingrum, and H. R. Hatta, "Machine vision for the maturity classification of oil palm fresh fruit bunches based on color and texture features," *Sci. Hortic. (Amsterdam).*, vol. 286, Aug. 2021, doi: 10.1016/j.scienta.2021.110245.

[23] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," *Computers, Materials and Continua*, vol. 80. Tech Science Press, pp. 1–35, 2024. doi: 10.32604/cmc.2024.053204.

[24] X. V. Lin *et al.*, "MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts," Jul. 2024, doi: 10.48550/arXiv.2407.21770.

[25] S. O. Ngesthi and L. A. Wulandhari, "Cassava Diseases Classification using EfficientNet Model with Imbalance Data Handling," *J. Online Inform.*, vol. 9, no. 2, pp. 148–158, Aug. 2024, doi: 10.15575/join.v9i2.1300.

[26] S. H. Shabiyya, B. H. Prasetio, and E. R. Widasari, "Harnessing the Power of CNN-Transformer Encoders in Stress Speech Analysis," in *Proceeding - International Conference on Information Technology and Computing 2023, ICITCOM 2023*, 2023, pp. 147–151. doi: 10.1109/ICITCOM60176.2023.10442454.