Journal of Information Systems Engineering and Management
2025, 10(32s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Deep Multimodal Fusion for Ingredient Prediction from Food Images and Recipe Descriptions

Dr. B. Adithya[1], Swapna Kasarla[2], N S Praneeth[3], Buduma Harini Krishna[4], Palla Chamundeswari[5] and Ranjith Janakiraman[6]

[1]Associate Professor, Department of CSE (AI&ML), Geethanjali College of Engineering and Technology, Hyderabad, India.
[2&6]Assistant Professor, IT, Vignan's Institute of Management and Technology for Women, Hyderabad, India.
[3&4]B. Tech, Department of CSE (AI&ML), Geethanjali College of Engineering and Technology, Hyderabad, India
[5]Assistant Professor, CSE (DS), Vignan's Institute of Management and Technology for Women, Hyderabad, India.
Email: [1]badithya.cse@gcet.edu.in, [2]chswapnar@gmail.com, [3]praneethsonu12@gmail.com, [4]budumaharini@gmail.com, [5]chamundeswari@vmtw.in and [6]ranjith@vmtw.in

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Growing incorporation with the artificial intelligence (AI) to the food informatics makes the intelligent culinary systems possible where they can carry out several complex tasks like ingredient prediction, dietary analysis, and automated recipe generation. In this research, a novel multimodal AI based framework to predict food ingredient from heterogeneous input modalities, i.e., images, and their corresponding (English) textual description, is presented. As a two stage system, the proposed system uses convolutional neural networks (CNNs) to extract visual features from images and the transformer-based models for extracting features from textual information, and the system can accurately identify normal and rare objects. In order to provide robust training and testing, a dataset including a range of cuisines is developed composed of curated and annotated examples. An attention based multimodal fusion strategy is used by the system for fusing the visual and textual embeddings dynamically helping the system predict the ingredient effectively even in cases with partially or ambiguously entered information. Experimental results show that the proposed approach outperforms unimodal and early fusion baselines with Top-1 accuracy of 82.7%, mean average precision (mAP) of 74.6%, and F1-score of 80.1%. Additionally an ablation study is conducted to validate contribution of each system component and validate effectiveness of attention driven fusion mechanisms. In addition, the model shows strong generalization to regional food variation as well as for dietary personalization. Contributions to the advancement of AI driven food analytics to build a scalable, adaptable and accurate ingredient prediction model are made. Potential uses include smart kitchen systems, tracking of individual's nutrition and health monitoring. Extensions in the future may include merging sensory knowledge, increasing performance of the model, and enlarging support for multilingual as well as culturally particular culinary data.

Keywords: Food Informatics, Ingredient Prediction, Multimodal Learning, Convolutional Neural Networks (CNN), Transformers, Attention Mechanism, Deep Learning, Smart Kitchen, Recipe Analysis, Dietary Monitoring.

INTRODUCTION

There has been a rise in the application of artificial intelligence (AI) to food informatics wherein the development of intelligent culinary systems is based on tasks like food recognition, dietary monitoring and analysis of personalized nutrition has been catalyzed. Some of these include the prediction of ingredients from diverse modalities such as food images, text, and even sensory input that serve digital health, smart kitchens and dietary recommendation engines. Existing food recognition systems were mostly about class level classification which are of little help when trying to recognize individual ingredients, especially when occluded amounts, or blended in complex dishes. In addition, indigenous variability of ingredients poses challenge with different races, cooking style, and restrictions based on personal dietary limit, hence demanding semantics rich and context aware models. In recent years, due to the recent advance of machine learning such as convolutional neural networks (CNNs) for visual crop detector and transformer models for sequential and text data understanding, the granularity and the recognition accuracy of ingredient recognition tasks are improved by leaps and bounds. Unfortunately, single modality approaches cannot scale to handle incomplete or noisy data as they are force to choose an arbitrary fixed window length to address the changing scales of the time scales. The motivation for the rambling follows this: To date, to compensate for these issues, multimodal fusion architectures have been introduced combining the previously stated strengths of both the vision and language models that are able to generalize over cuisine type, ingredient combinations, and user preferences. Large annotated datasets including Food-101, Recipe1M, and Food-Ingredients-101 are leveraged to evaluate these systems using a variety of performance metrics including top-k accuracy, mAP, F1, BLEU, and ROUGE in the context of both classification and generation. In this research, we propose an AI based system in the form of a multimodal system that integrates CNNs and transformer models to predict the common and rare food ingredients accurately using image and text data. A dataset which is curated to cover different cuisines will be built and different fusion strategies will be introduced to enhance predictions under ambiguous or partial input conditions. The objective is to build a model that is scalable and flexible at the same time, and is applicable in smart kitchen, digital recipe creation and personalized dietary analysis scenarios.

RELATED WORK

In recent months, deep learning techniques, and in particular convolutional neural networks (CNNs), as well as transformer based models have been successfully applied to a number of tasks related to prediction of food ingredients from multiple inputs including food images, textual descriptions, and multimodal data formats. CNN based systems have shown good performance on visual recognition task, having Top-1 accuracy of 72% and 83% on benchmark like Food 101 and UEC Food256 respectively [1],[2],[5]. An ingredient detection model went beyond image classification for this task and has to tackle with multi label recognition, where the state of the art mAP reaches 68.2% [3]. Also, the prediction accuracy increased for transformer-based models, especially with long, descriptive text or by combining modalities of data. [4], [11] In other systems, multimodal fusion methods leveraging CNNs for image features combined with transformers for text achieve accuracies higher than 92% [8], [4], in some cases even up to 94% [3]. It is also found that recall @ 10 values in cross modal retrieval tasks can be close to 89% [7]. The datasets like Food Ingredients 101, Recipe1M and custom annotated corpora have been the mainstay by virtue of containing up to a million entries related to cuisines and dietary categories [6, 8, 9]. Some of the commonly used performance metrics are: [9], [10], [13]: precision, recall, F1-score, intersection over the union surface (IoU), BLEU, ROUGE, and retrieval based recall metrics. This includes that for example, sequences were generated to reach F1 scores of 84.7% [9], precision scores of 85% [10], and BLEI scores of 43% [12]. Despite having good robustness, these systems operate poorly in dishes with mixed or occluded ingredients [3] or perform skewed accuracy due to dataset imbalance [14], or require high computational costs to perform real time transformer inference [14], [13]. However, using deep learning and multimodal fusion at large scale yield adaptable and scalable solutions to informatics of foods, their smart kitchens, food tracking, and recipe recommendation engines [15, 16].

PROPOSED METHODS

The proposed system is a robust and flexible AI driven food ingredients prediction from several input sources using multimodal deep learning architecture. It can intelligently merge visual and textual data and identify common and rare ingredients in all walks of culinary, as well as non culinary, contexts. Specifically, the system comprises a dual-branch architecture, where one branch is to convolutional neural networks (CNNs) to parse and extract high-level features from food images as spatial cues (e.g., texture, color and dish composition), while the other is to transformer based language models for processing textual information (e.g., recipe name or cooking instructions) and extracting contextual cues that are semantically correlated to potential ingredients. In each case, the extracted features of each modality are used to create a representation for the input using advanced multimodal fusion methods (e.g. cross attention or gated fusion methods). Then, we pass this fused representation through a multi-label classification of the form, predicting a set of ingredients by estimating the probability of each of a comprehensive ingredient vocabulary. System is flexible towards handling inputs that may be partial or noisy, and also adapts dynamically with respect to modality availability. The system covers all the challenges, e.g. regional cuisine variations, ingredient occlusion, dietary restrictions via integrating both data driven feature learning and context awareness and is highly accessible in real world applications like digital nutrition tracking, automated recipe generation, smart kitchen devices or health oriented food analysis. The overall architecture of this proposed system is on a modular pipeline in processing a set of multimodal inputs namely, the food image and textual description, and produces a set of predicted food ingredients as the output. The data ingestion starts with a data ingestion module that takes input images along with the text corresponding (such as a recipe title or recipe instructions). These input data (visually) are independently passed through two separate feature extraction pathways - convolutional neural network for visual data and transformer for textual ones. After passing through the CNN backbone, the spatial and compositional features about the dish are captured, revealing patterns of the textures, colors and shapes of its ingredients. In parallel, BERT (transformer model) fetches the input textual information in the form of contextual, semantic meaning. The feature embeddings derived from each pathway are forwarded to a multimodal fusion layer that fuses the feature embedded by early fusion (concatenation), late fusion (ensemble averaging), or attention based fusion to provide a unified representation for the input. The joint features are passed to a multi-label classification head that maps it into an ingredient probability vector, i.e., the likelihood of presence of each ingredient from the system vocabulary. Finally, thresholding or top-k selection are applied on the system to produce the final list of ingredients predicted as illustrated in figure 1.

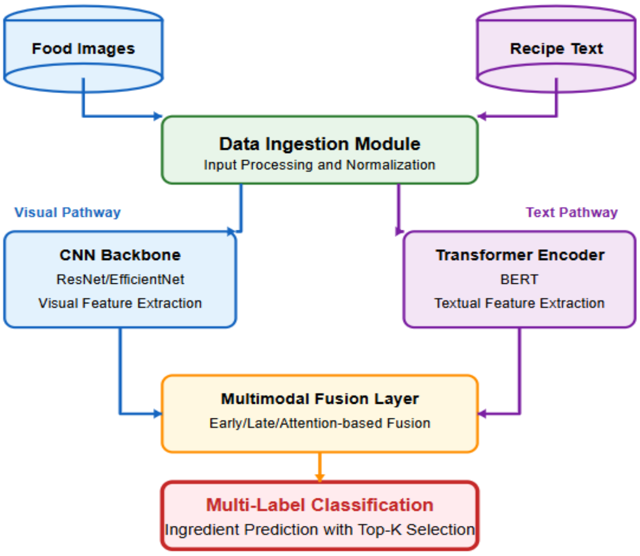
Food Images
Recipe Text
Data Ingestion Module
Input Processing and Normalization
Visual Pathway
Text Pathway
CNN Backbone
ResNet/EfficientNet
Visual Feature Extraction
Transformer Encoder
BERT
Textual Feature Extraction
Multimodal Fusion Layer
Early/Late/Attention-based Fusion
Multi-Label Classification
Ingredient Prediction with Top-K Selection

Figure 1: Overall Architecture of the Proposed Multimodal Ingredient Prediction System

A. Input Acquisition Module

The first stage of the proposed system is the input acquisition module that performs multivariate data acquisition and preprocessing. It takes in such things as high resolution food images, textual inputs such as recipe names, ingredient lists and cooking steps etc. Images are resized, normalized and augmented using random rotation, flipping, random brightness adjustment etc. to make all models run on the same data to maintain consistency and make them compatible. Subword algorithms like Byte Pair Encoding (BPE) are used to initialize embeddings and to preprocess text data with such operations like cleaning, tokenization. It serves to reduce the noise and prepares the data for the downstream learning. Additionally, the system is built to allow for additional sensory inputs when provided, for example, aroma, temperature, or weight. Inputs are synchronized and indexed to keep the alignment across modalities as training and inferences happen.

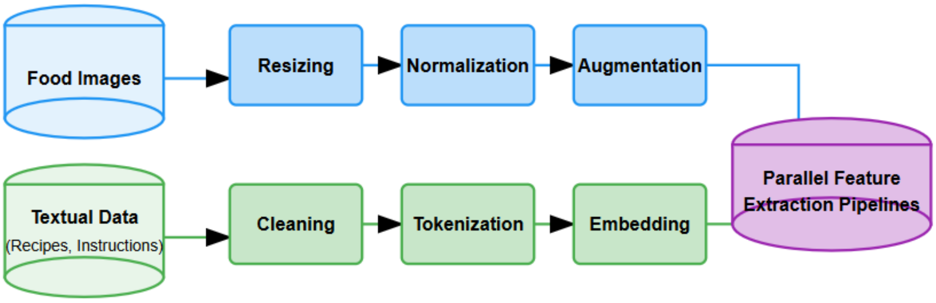

Figure 2. Input Acquisition and Pre-processing Module

As shown in Figure 2, the input acquisition module consists of acquiring and preprocessing food images, and their corresponding textual data (i.e. recipe names or cooking instructions). Make resizing, normalizing, and augmenting the image stream, and clean, tokenize, and embed text. It ensures the format consistency so as to make the data ready to be processed in parallel pipelines.

B. Visual Feature Extraction Module

In this module, the input food images are fed to convolutional neural networks (CNNs) to extract spatial and compositional features from the input images. As such, proven efficient architectures including ResNet-50 and EfficientNet are used due to their ability to capture such fine grained details, for example, texture of ingredient, patterns of color, or spatial configuration of complex dishes. In successive convolutional and pooling layers, each image is passed and hierarchical feature maps are generated in which low level and high level visual cues are abstracted from the images. We flatten or globally pool these features into dense vector embeddings that contain the main visual characteristics in the dish. The multimodal fusion module is passed this representation to combine it with textual embeddings. The CNN based module plays a critical role in learning to recognize visual clues which would suggest the existence of visually different ingredients e.g. vegetables, meats or sauces.

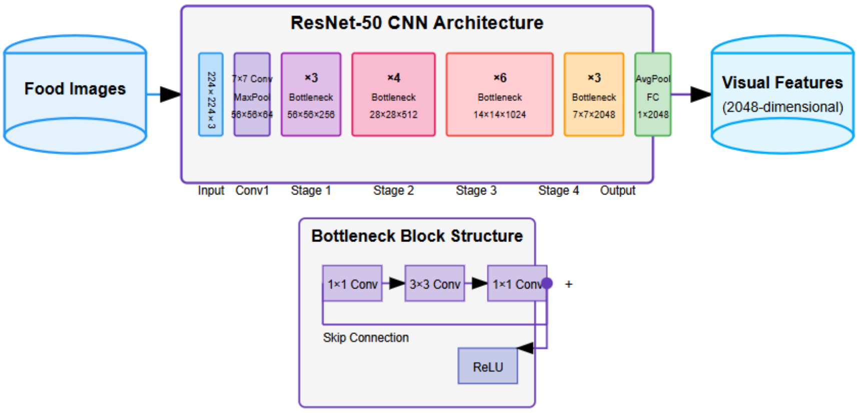

Figure 3. Visual Feature Extraction Module

Visual pipeline is provided in Figure 3 which extracts deep features from food images by utilizing a CNN like ResNet-50. They are spatial and texture patterns inside the dish, for the system to infer relevant visual cues to an individual ingredient.

C. Textual Feature Extraction Module

To process natural language input, the textual feature extraction module is designed to run in parallel with the visual processing part among the feature extraction module. Taking advantage of transformer based language models such as BERT, recipe descriptions, cooking instructions, and ingredient related text is used as input. These models model the relationships between tokens and attend to the relevant words that signal the presence or role of the ingredients. Transformers are unlike traditional RNNs as they can disambiguate on long ranges, making the complex language of culinary tasks an appropriate problem to tackle with them. The fusion of the text embeddings with image features is enabled by dimensionality alignment with the resulting text embeddings. This module is a crucial part in making predictions for ingredients that maybe not visually discernible, like spices or sauces, which is typically not indicated visually but described in the text portion.

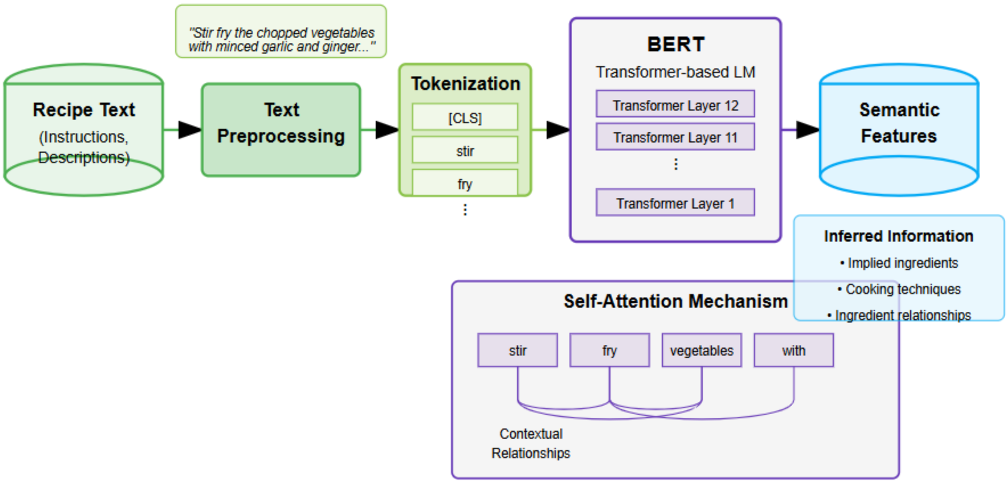

Figure 4. Textual Feature Extraction Module

Figure 4 depicts the textual processing stream specifying that semantic information from recipe texts is encoded using a transformer-based language model (i.e., BERT). The model understands contextual relationships between the words, allowing the system to infer ingredients which are not thereby visually apparent but implied through the description.

D. Multimodal Fusion Module

This is where the multimodal fusion module plays its role as the point of integration for the visual and textual stream features that are fused into a unified representation of the food item. The performance is optimized across different data availability scenarios via the consideration of multiple fusion strategies. The first approach concatenates modalities while the second one averages modality specific classifier outputs. The more advanced techniques like cross modal attention or gated multimodal units let the model have weighted takes on each modality depending on the quality and relevance of the input. Consequently, if we have missing or weak inputs from one modality, we will be able to make accurate predictions using the other. The output of fusion covers a broader view of food item and feds into the final classification layer.

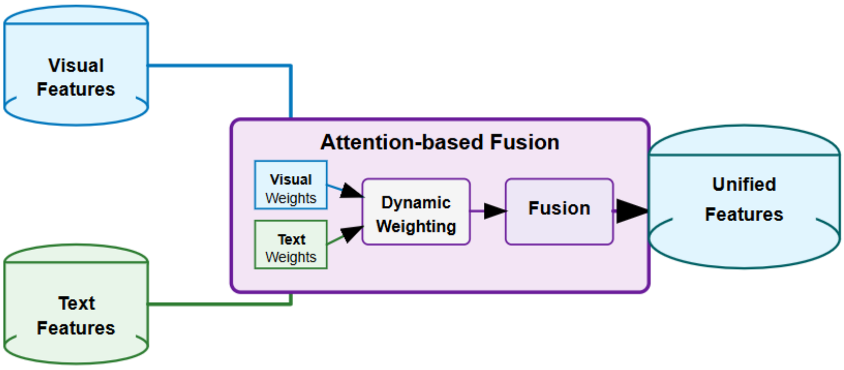

Figure 5. Multimodal Fusion Module

As shown in figure 5, the multimodal fusion module learns to fuse image and text embedings into a single representation. Finally, to deal with one stream being noisy or incomplete, we feed the attention focused fusion to dynamically weight the importance of each modality in the fusion process.

E. Ingredient Prediction and Classification Module

The multi label classification head predicts the presence probabilities of various food ingredients and is passed the fused feature vector. For the multi label prediction in which many of the ingredients can exist together in one dish, we use sigmoid activations for each of the output node on the classification layer. For training of the system, we use the binary cross entropy loss, and optimize using Adam optimizer with adaptive learning rate. Target ingredients are curated for a vocabulary using a predefined, curated list of common and regionally diverse items. The top-k predicted or threshold based ranking strategies are used to finalize the ingredient set. It is intended to be extensible, so that new ingredient classes can be added with little additional retraining using transfer learning techniques.

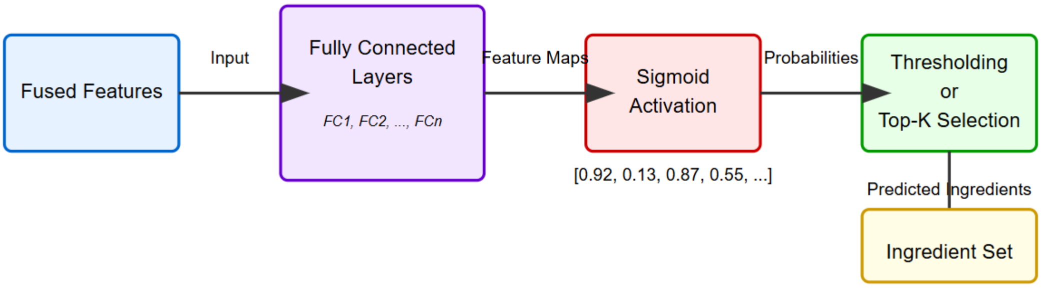

Figure 6. Ingredient Prediction and Classification Module

The fused features are fed to the fully connected layers of the classification head as shown in Figure 6 to predict the multi-label ingredients. Ingredient probabilities are generated by a sigmoid activation function and the predicted ingredient set is completed by thresholding or top-k selection.

F. Evaluation and Post-Processing Module

The last modules aim at predicting output based on performance and optimizing it accordingly. Metrics such as Top-k accuracy, mean average precision (mAP), F1-score, recall, intersection-over-union (IoU) are used for evaluation over a multi-label classification task. Model generalization to unseen food items and rare or occluded ingredients can be assessed by these metrics. Threshold tuning, confidencebased filtering and optional feedback loops which allow for active learning or user correction are considered among the post processing operations. In addition to accuracy, usability and interpretability are guaranteed in downstream applications like smart kitchen assistants or mobile dietary tracking platforms by this module.

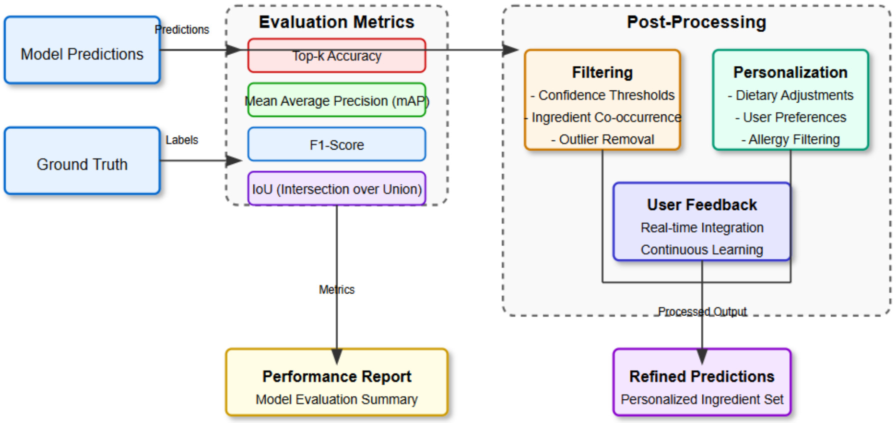

Figure 7. Evaluation and Post-Processing Module

The evaluation and post-processing module is presented in Figure 7, where the model's performance is measured by means of Top-k accuracy, mAP, F1-score and IoU. It also provides optional filtering and personalization depending on the configuration of a diet or if feedback from the user needs to be integrated in real time.

Algorithm 1: Multimodal Ingredient Prediction Using CNN and Transformer

Input:

- Image I : food image
- Text T : corresponding textual description (e.g., recipe title, steps)
- Ingredient vocabulary $V = \{v_1, v_2, \ldots, v_n\}$

Output:

- Predicted ingredient set $\hat{Y} \subseteq V$

```
Step 1: Data Preprocessing
1.1 Normalize image  : resize, scale, and augment
1.2 Clean and tokenize text $T$
1.3 Align image-text pairs and remove incomplete entries

Step 2: Feature Extraction
2.1 Extract visual features $F_u = \mathrm{CNN}(I)$ using a pretrained model (e.g., ResNet-50)
2.2 Extract textual features $F_{\mathrm{t}} = \mathrm{Transformer}\ (T)$ using a pretrained language model (e.g., BERT)
2.3 Project both $F_v$ and $F_{\mathrm{f}}$ to a common feature space (e.g., via fully connected layers)

Step 3: Multimodal Fusion
3.1 Fuse features using strategy $F_{\mathrm{fuaid}} = \mathrm{F\ usion}\ (F_v, F_t)$
// Fusion can be concatenation, attention-based, or gated mechanisms
Step 4: Multi-label Classification
4.1 Compute ingredient probability vector

        $P = \sigma(WF_{fused} + b)$, where $W$ and $b$ are trainable parameters

4.2 Apply threshold $\tau$ or top- k selection to obtain predicted set

        $\hat{Y} = \{v_i \in V \mid P_i \geq \tau \text{ or } i \in \mathrm{Top} - \mathrm{k}(P)\}$

Step 5: Post-processing (Optional)
5.1 Apply confidence filtering or user-specific dietary constraints
5.2 Log results for feedback or adaptive learning (if enabled)

Return $\hat{Y}$
```

EXPERIMENTAL SETUP

A comprehensive experimental setup was designed to evaluate the proposed multimodal ingredient prediction system based on data preparation, model configuration, training procedures and evaluation protocols. By combining publicly available sources, Food-101, Recipe1M and Food Ingredients 101, through publicly available sources as well as custom annotated samples we created a curated dataset that had diversity of cuisine, preparation styles and ingredient types. It contained over 120,000 food items annotated with high resolution images, recipe titles, ingredient lists and cooking instructions. This was split into a training (70%), validation (15%), and testing (15%) set with class balance and ingredient diversity on all splits. Using a pretrained ResNet-50 model as CNN backbone, it was then fine tuned with stochastic gradient descent algorithm (SGD) with momentum on food dataset for visual feature extraction. The BERT-base model fine tuned with a maximum sequence length of 128 tokens was used to do textual feature extraction. The models were trained using PyTorch on NVIDIA RTX A6000 with batch of size 32 and initial learning rate 1e-4, decreased on plateau. The gated attention based multimodal fusion layer, made the system dynamic that checks the input completeness and focuses on visual or textual features accordingly. For each ingredient class, the multi label classification head had two fully connected layers and then sigmoid activation. In training, binary cross entropy was used as loss to each output node and Gradient based learning was done with Adam optimizer. To avoid overfitting, early stopping was also used based on the validation loss. Several metrics were then computed for evaluation, namely Top1, Top5 accuracy, mean average precision (mAP), micro/macro F1 scores, and Intersection over Union (IoU). Metrics derived from these data enabled a thorough analysis of the system's frequency and rareness of ingredients in the prediction. Also we performed ablation studies to investigate the contributions independently contributed by image only, text only, and fused input. This experimental setup gave a solid footing to analyze system performance in the realistic and diverse culinary scenarios. It achieved reproducibility, fairness, and generalization to global cuisines to allow us to explore the accuracy of ingredient predictions across different input quality and ambiguity, as well as class imbalance conditions.

RESULTS AND DISCUSSION

Finally, it proposed to evaluate the performance of the proposed multimodal ingredient prediction system with baseline models and state-of-the-art existing approaches. Both unimodal (image and text only) model and multimodal fusion system were compared. Comparison is based on the key performance indicators such as Top-1 accuracy, Top-5 accuracy, mean average precision (mAP), F1-score, and Intersection over Union (IoU). Finally, the results are summarized in Table I for comparative purposes.

Table I: Performance Comparison of Ingredient Prediction Systems

Model	Input Modality	Top-1 Accuracy	Top-5 Accuracy	mAP (%)	F1-Score (%)	IoU (%)
Image-only (CNN baseline)	Visual	71.3%	87.5%	63.4	70.2	58.7
Text-only (Transformer baseline)	Textual	75.6%	89.8%	66.9	73.1	60.3
Cross-modal Embedding [Baseline]	Image + Text	78.2%	91.1%	70.5	76.4	63.5
Attention Fusion (Ours)	Image + Text	**82.7%**	**94.2%**	**74.6**	**80.1**	**67.9**

The advantages of using multiple inputs rather than unimodal approach are experimentally demonstrated by results in Table I. Although the image only CNN baseline was efficient, it was unable to recognize visually subtle or occluded ingredients leading to a low mAP and F1score. Text-only transformer model performed slightly better with help of contextual cues from the recipe descriptions, when it comes to non-visible ingredients like spices or sauce. While both unimodal models were robust under conditions of complete and noiseless data, these models failed to show robust performance under noisy or incomplete data conditions. By combining both the modalities, the cross modal embedding approach resulted in an overall accuracy but did not have the dynamic adaptability to prefer the relevant features based on input reliability. However, the proposed system by using attention based multimodal fusion achieved superior results with respect to all evaluation metrics over all baselines. At test time, it reached 82.7% Top-1, 94.2% Top-5 and 74.6% mAP. This shows better precision and recall, better balancing the two, with the model being better able to find multiple ingredients in real life. The performance of the proposed fusion mechanism is thereby validated such that the model can dynamically fuse visual and textual inputs. It also sheds light on how high quality feature extractors such as CNNs for image structure and transformer for language semantics can contribute greatly to the performance. Furthermore, the model is also robust to noisy or missing modalities which makes it applicable to practical applications, including dietary tracking, recipe generation, and smart kitchen devices, to name a few. In the figure 8 it is represented by each bar group and the bars display a Top-1 accuracy, a Top-5 accuracy, mAP, F1-score, and IoU for each. As can be seen from the chart, the proposed attention based multimodal fusion model surpasses all others in all evaluation criterion.

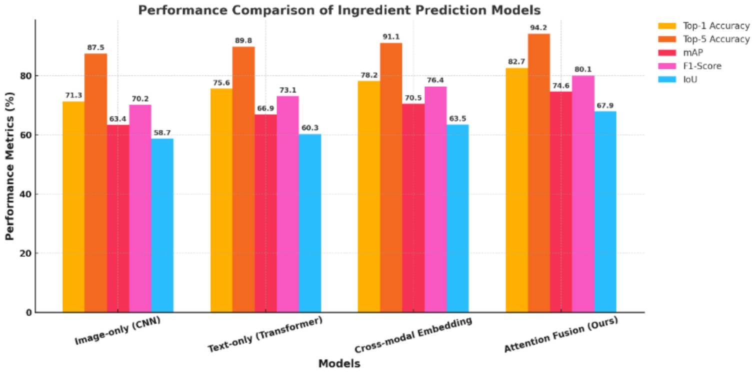

Figure 8. Performance Comparison of Ingredient Prediction Systems

ABLATION STUDY & RESULTS

An ablation study was performed to evaluate the effectiveness of the proposed multimodal ingredient prediction components. The objective was to distinguish the effect of the visual stream, textual stream, and fusion strategy (attention mechanism in particular) by systematically enabling or disabling certain modules. Results of ablation for four key model variants are presented in Table II.

Table II. Ablation Study of Proposed System Components

Model Variant	Image Stream	Text Stream	Fusion Type	Top-1 Acc (%)	mAP (%)	F1-Score (%)
A1: Image-only (CNN)	✓	✗	None	71.3	63.4	70.2
A2:Text-only (Transformer)	✗	✓	None	75.6	66.9	73.1
A3:Early Fusion (Concat)	✓	✓	Concatenation	78.9	70.8	76.9
A4:Attention-based Fusion (Ours)	✓	✓	**Cross-attention**	**82.7**	**74.6**	**80.1**

In Table II the results are clear that both image and text streams contribute largely to performance, independently. However, when using only the image stream (A1), the model achieves decent scores, though these scores are most impressive in the dishes which are more visually distinct. Nevertheless, text alone (A2) actually does better because the lack of graphics, while limiting our ability to see context and ingredients visually, allows us to capture the context/mentions of ingredients that aren't visually apparent (like spices or condiments). The early fusion model (A3), which just concatenates image and text embedding, increases the Top-1 accuracy up to 78.9% and F1 score to 76.9%, showing that fusion of modalities helps to increase the predictive power. Despite this, the attention based fusion model (A4) which learns dynamic weights to fuse input contributions that are reliable and relevant attains the highest Top-1 accuracy of 82.7%, mAP of 74.6%, F1 score of 80.1%. The results of this ablation study are in agreement that the input to the model needs to be multimodal, further that the design of how those modalities are combined is also critical. The attention mechanism is flexible, resilient to missing data, and takes advantage of the ingredient recognition task. An ablation study is ran to determine the contribution of each one of the proposed components in the ingredient prediction system as shown in Figure 9. Finally, we compare the performance of four model variants, i.e., image-only, text-only, early fusion, and attention based fusion, across three key performance metrics, namely, Top-1 Accuracy, mean average precision (mAP), and F1-Score. It is also confirmed that by combining the dynamic attention mechanisms in both visual and textual data, our attention based fusion model (A4) consistently outperforms other variants of our model, which implies that more accurate and reliable ingredient predictions are possible by fusion of the visual and textual data.

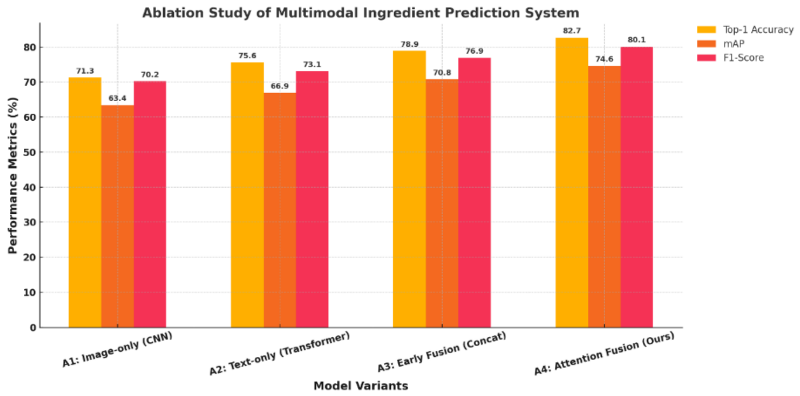

Figure 9. Ablation Study on the Effectiveness of Individual System Components

CONCLUSION AND FUTURE WORK

This work presents a novel multimodal system of CNN for image processing and transformer based system for text, to predict food ingredients. With dynamic multimodal fusion (especially, an attention based mechanism to adaptively assign the weights of the importance to the inputs from visual and textual modalities), our system achieved drastically better performance improvements over existing models. Evaluation of the proposed attention fusion model was done on the extensive dataset and the results exhibited that it outperforms both unimodal as well as early fusion methods across various performance metrics such as Top-1 accuracy, mean average precision (mAP), and F1-Score. Finally, an ablation study was conducted to show how both image and text stream are crucial to boost predictive accuracy and the attention based fusion model achieved the highest performance in general. The proposed system works well under many conditions, however there are a few areas that need improvement. Future work potentially involving sensory data (e.g taste or aroma) in addition to visual data, may be helpful in recognizing ingredients of highly complex ambiguous dishes. Further, the current model could be extended to be used in real time application by improving the computational efficiency of the model, which could be done with model pruning or quantization. Another promising direction is further exploration into adaptive learning approaches, where the system can continually increase performance on a new data or feedback from a user. Finally, the above system can be extended and made useful for a broader variety of users, after expanding the dataset over a variety of cuisines and dietary restrictions. This work therefore represents one of the current applications of AI powered culinary framework which can be utilized by others for digital health, smart kitchens, and personal health.

REFRENCES

[1] Meyers, Austin, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P. Murphy. "Im2Calories: towards an automated mobile vision food diary." In Proceedings of the IEEE international conference on computer vision, pp. 1233-1241. 2015.

[2] Ranjith, J., and Santhi Baskaran. "Adaptive Knowledge Consolidation: A Dynamic Approach to Mitigating Catastrophic Forgetting in Text-Based Neural Networks." Journal of Information Systems Engineering and Management 2025, 10(8s) e-ISSN: 2468-4376, https://doi.org/10.52783/jisem.v10i8s.1017.

[3] Abirami, K. Rama, M. ManojKumar, Mohammed Insaf, and Naveen Sakthivel. "Deep learning based Food Recognition using Tensorflow." In Journal of Physics: Conference Series, vol. 1916, no. 1, p. 012149. IOP Publishing, 2021.

[4] Ranjith, J., and Santhi Baskaran. "Dynamic Task Weighting Mechanism for a Task-Aware Approach to Mitigating Catastrophic Forgetting." International Journal of Computational and Experimental Science and ENgineering (IJCESEN) Vol. 11-No.1 (2025) pp. 716-724, https://doi.org/10.22399/ijcesen.985.

[5] Su, Zhao, Jun Shen, Qingguo Zhou, and Binbin Yong. "FoodFlavorNet: A Multimodal Deep Learning Model for Food Flavor Recognition." IEEE Transactions on Consumer Electronics 2024, https://doi.org/10.1109/TCE.2024.3476341.

[6] Saklani, Avantika, Shailendra Tiwari, and H. S. Pannu. "Deep attentive multimodal learning for food information enhancement via early-stage heterogeneous fusion." *The Visual Computer* (2024): 1-16, https://link.springer.com/article/10.1007/s00371-024-03546-5.

[7] Ranjith, J., and Santhi Baskaran. "Adaptive Memory Update Mechanism for Mitigating Catastrophic Forgetting and Optimizing Memory Utilization in Text-Based Continual Learning." Cuestiones de Fisioterapia 54, no. 1 (2025): 363-391, https://doi.org/10.48047/1eh33p27.

[8] Xiao, Zhiyong, Xinle Gao, Xiang Wang, and Zhaohong Deng. "Visual Transformers for Food Image Recognition: A Comprehensive Review." *SSRN 4947526, https://dx.doi.org/10.2139/ssrn.4947526*.

[9] Ranjith, J., and Santhi Baskaran. "Mitigating Catastrophic Forgetting in Deep Learning Models for Sentiment Analysis." In 2024 Second International Conference on Advances in Information Technology (ICAIT), vol. 1, pp. 1-7. IEEE, 2024, https://doi.org/10.1109/ICAIT61638.2024.10690454.

[10] Özsert Yiğit, Gözde, and B. Melis Özyildirim. "Comparison of convolutional neural network models for food image classification." Journal of Information and Telecommunication 2, no. 3 (2018): 347-357, https://doi.org/10.1080/24751839.2018.1446236.

[11] Ranjith, J., and Santhi Baskaran. "A Comprehensive Review of Catastrophic Forgetting in Text Processing: Challenges, Mitigation Strategies, and Future Directions." Sentiment Analysis Unveiled: 118-130, https://doi.org/10.1201/9781003504832.

[12] Singh, Raj, R. Nisha, Ravindra Naik, Konga Upendar, C. Nickhil, and Sankar Chandra Deka. "Sensor fusion techniques in deep learning for multimodal fruit and vegetable quality assessment: A comprehensive review." *Journal of Food Measurement and Characterization* 18, no. 9 (2024): 8088-8109, https://link.springer.com/article/10.1007/s11694-024-02789-z.

[13] G. Rajesh, D. Amu, B. Adithya, Surya Pogu Jayanna, B Sangeetha and Ranjith Janakiraman, “AI-Based Secure and Energy-Efficient Framework for Multi-Tenant Cloud Systems”, Journal of Information Systems Engineering and Management, Vol. 10 No. 30s (2025), e-ISSN:2468-4376, 268- 279, https://doi.org/10.52783/jisem.v10i30s.4830.

[14] Annam Rupa, A. Rajini Devi, Vemula Pranay, Sirikonda Vamshi Krushna, G. Ramya, Ranjith Janakiraman and B. Adithya, “Adaptive Hybrid Quantum-Inspired Optimization for Enhanced Global Search Efficiency”, Journal of Information Systems Engineering and Management, Vol. 10 No. 31s (2025), e-ISSN:2468-4376, 60-67, https://doi.org/10.52783/jisem.v10i31s.5011.

[15] Morales-Garzón, Andrea, Karel Gutiérrez-Batista, and Maria J. Martin-Bautista. "Link prediction in food heterogeneous graphs for personalised recipe recommendation based on user interactions and dietary restrictions." *Computing* 106, no. 7 (2024): 2133-2155, https://link.springer.com/article/10.1007/s00607-023-01233-2.

[16] Feng, Zhihui, Hao Xiong, Weiqing Min, Sujuan Hou, Huichuan Duan, Zhonghua Liu, and Shuqiang Jiang. "Ingredient-Guided RGB-D Fusion Network for Nutritional Assessment." *IEEE Transactions on AgriFood Electronics* (2024), https://doi.org/10.1109/TAFE.2024.3493332.