A Comparative Study of Early Fusion and Multimodal Siamese Neural Network in Food Classification

Kanokporn Sintarasirikulchai, Akarin Phaibulpanich and Seksan Kiatsupaibul

January 8, 2025

A Comparative Study of Early Fusion and Multimodal Siamese Neural Network in Food Classification

Kanokporn Sintarasirikulchai[1*], Akarin Phaibulpanich[2], and Seksan Kiatsupaibul[3]
[1]Department of Statistics, Chulalongkorn University, Bangkok, 10330, Thailand, 6580232426@student.chula.ac.th
[2]Department of Statistics, Chulalongkorn University, Bangkok, 10330, Thailand, akarin@cbs.chula.ac.th
[3]Department of Statistics, Chulalongkorn University, Bangkok, 10330, Thailand, seksan@cbs.chula.ac.th
(* corresponding author)

Abstract

The shift from traditional diets to those high in fat and calories has led to an increase in obesity and related health issues, emphasizing the need for accurate dietary monitoring. Automated systems utilizing artificial intelligence (AI) have emerged as promising tools for providing personalized dietary advice through food classification. With the growing volume of food-related content on social media, including images and accompanying text, there is an increasing need to leverage multimodal data for more accurate predictions. This study compares two fusion techniques for food classification: early fusion and the multimodal Siamese Neural Network (mSNN). Using the UPMCFood-101 dataset, which includes images and text descriptions across 101 food categories, the analysis focuses on three specific classes: bread pudding, chicken wings, and waffles. While the study focuses on these three classes, the flexible architecture of both models suggests their potential for generalization to other food categories. The early fusion model demonstrated strong generalization, achieving an overall accuracy of 0.960. In contrast, the mSNN, trained with 72,000 pairs, achieved a peak accuracy of 0.976, outperforming the early fusion model in precision, recall, and accuracy, particularly with smaller image text per class in databases. However, the mSNN's performance declined with larger databases due to outlier effects that skewed average distance calculations, leading to reduced accuracy. These findings suggest that while the mSNN is more accurate with smaller datasets, the early fusion model provides better generalization.

Keywords: Multimodal, Early fusion, Multimodal Siamese Neural Network

1. Introduction

Food plays a vital role in maintaining human health by providing the necessary energy and nutrients. However, dietary habits have significantly shifted with economic growth, moving from traditional diets to those high in fat, calories, and low in fiber. This transition has been linked to the growing prevalence of obesity and related health issues. Consequently, the importance of monitoring dietary intake and ensuring balanced nutrition has become increasingly recognized. To address these challenges, there is a growing interest in leveraging automated systems that provide personalized dietary advice, promoting healthier eating habits and reducing the risks associated with poor nutrition. In this context, artificial intelligence (AI) has emerged as a key tool in developing methods for accurately classifying food images, which is essential for effective nutritional management [1,2].

Recent advances in AI, particularly in deep learning, have significantly improved the accuracy of food classification. Convolutional Neural Network (CNNs) have been particularly successful in these tasks. The widespread use of smartphones has led to a surge in food images and related text shared on social media [3]. Consequently, vast amounts of food-related data, encompassing both images and text, have become readily available, presenting new opportunities and challenges for multimodal data fusion. The integration of these diverse data modalities has become a critical area of research, enabling models to leverage complementary information from various sources, thereby improving performance in complex classification tasks.

This study focuses on comparing two fusion techniques—early fusion and multimodal Siamese Neural Network (mSNN) —in the context of food classification. Fusion techniques are crucial for enhancing model accuracy and robustness by integrating different types of data, such as images and textual information. The UPMCFood-101 dataset, a comprehensive collection of images paired with text descriptions across 101 food categories, serves as the foundation for this research. Due to computational constraints, this study focuses on three specific classes from the dataset—bread pudding, chicken wings, and waffles—selected based on their high recall in preliminary testing with the early fusion model. In prior work, Gallo et al. employed early fusion on the full dataset, achieving high accuracy and demonstrating the effectiveness of this approach for food classification [4]. However, with the development of more advanced methods like multimodal Siamese network, it is necessary to explore whether these newer techniques can further enhance performance.

Although this study specifically tests three food classes, the model architecture developed here can be generalized and applied to a broader range of food categories with minimal or no architectural modifications. The methodology used, which combines both image and text data, is designed to handle a variety of classes without requiring model-specific adjustments for individual categories. However, it is important to note that prediction accuracy may vary depending on the visual characteristics and textual descriptions specific to each food type. The flexibility of this multimodal data fusion approach suggests that it can be extended to include additional food types beyond those used in this experiment. Nevertheless, considerations must be given to computational requirements, as increasing the number of food categories directly impacts both training time and computational complexity. Therefore, while this work focuses on a limited number of classes due to computational constraints, the findings are expected to be applicable to other food categories, making the model suitable for real-world applications.

Multimodal Siamese Neural Network (mSNN), introduced by Chakladar et al., offers a novel approach by learning discriminative features from different data modalities and integrating them into a common feature space [5]. Originally designed for biometric verification, this method has proven highly effective in tasks requiring the fusion of spatial and temporal features, such as image and EEG signal analysis. Given its success in these domains, it is essential to evaluate its potential for food classification, particularly in comparison to traditional early fusion techniques.

The motivation for this research lies in identifying the most effective fusion technique for food classification, a task with significant implications for health monitoring and dietary analysis. By comparing early fusion with multimodal Siamese fusion using the UPMCFood-101 dataset, this study aims to determine which method offers superior performance in terms of accuracy and robustness.

This paper is structured as follows: Section 2 details the research methodology, including the dataset, model architectures, and evaluation metrics. Section 3 presents the research results and discussion. Finally, the conclusion in Section 4 summarizes the key findings of this study and suggests directions for future research.

2. Research Methodology

The following sections describe the dataset used in this study, the preprocessing steps applied to the data, the implementation of early fusion and multimodal Siamese fusion techniques, and the metrics used for evaluation.

2.1 Dataset

The UPMC Food-101 dataset, sourced from the Kaggle platform [4], is utilized in this study. Due to computational resource constraints, this research focuses on the three classes with the highest recall when tested using the early fusion model: bread pudding, chicken wings, and waffles.

These high-recall classes were selected to ensure that the models are evaluated on categories with reliable data. Table 1 provides a detailed breakdown of the dataset composition, including the number of image-text pairs for each class in the training, validation, and test sets. Table 2 presents examples of the image and text pairs for each class.

Table 1: Summary of dataset composition

Class	Training dataset	Validation dataset	Test dataset
Bread_pudding	610	67	226
Chicken_wings	590	65	219
Waffles	600	70	235
Total	1800	202	680

Table 2: Examples of image and text pairs for each food category

Class	Image	Text	Class	Image	Text
Bread pudding		Bread Pudding II Recipe - Allrecipes.com	Bread pudding		POLISHING OFF. . .: TOFFEE BREAD PUDDING W/ CINNAMON TOFFEE SAUCE
Chicken wings		Epic Dry-Rubbed Baked Chicken Wings - The Chunky Chef	Chicken wings		Chicken Wings with Blue Cheese Dip Recipe | MyRecipes.com
Waffles		Breads and Doughs-Pancakes and Waffles Recipes - Fine Cooking	Waffles		5 Quick Breakfast Recipes for Kids | Fashion Blog - Fashionandyou.com

2.2 Feature Extraction

2.2.1 Image encoder

The preprocessing of images involves resizing them to dimensions of 299 x 299 pixels, followed by normalization to a scale of 0 to 1. After preprocessing, the images are input into the InceptionV3 model with the last two layers removed. Subsequently, average pooling with a filter size of 8 x 8 is applied, followed by dropout with a probability of 40%, and then flattening is performed. Finally, a dense layer is used to reduce the dimensionality to 128, resulting in extracted image features. This architecture is based on the work by Gallo et al. [4] and is illustrated in Figure 1.

The input image is defined as i, the image model defined as $\mathcal{H}$ and the output of $\mathcal{H}$ is given as:

$$Y_{\mathcal{H}} = \mathcal{H}\ (i) \tag{1}$$

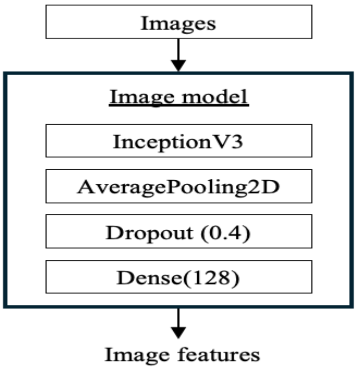

Figure 1: Diagram of the image encoder

2.2.2 Text encoder

In this research, several techniques are applied to preprocess the text data, including converting all text to lowercase, removing punctuation and numbers, eliminating single characters, and reducing multiple spaces. The BERT model is configured with the following parameters: English language uncased, 12 hidden layers, 768 hidden units, 12 self-attention heads, a vocabulary size of 30,522 words, and a total of 110 million parameters. The preprocessed text is input into the BERT model, followed by a Long Short Term Memory networks (LSTM) layer with 128 units, resulting in extracted text features. This approach follows the methodology described by Gallo et al. [4] and is illustrated in Figure 2.

The input text is defined as t, the text model defined as $\mathcal{W}$ and the output of $\mathcal{W}$ is given as

$$Y_{\mathcal{W}} = \mathcal{W}(t) \tag{2}$$

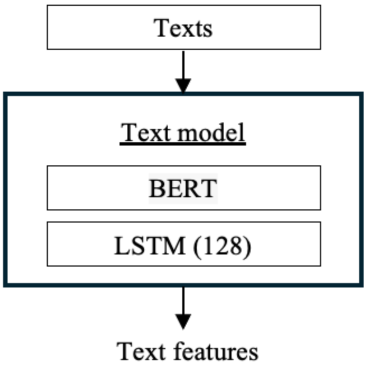

Figure 2: Diagram of the text encoder

2.3 Data fusion

2.3.1 Early fusion

The architecture of early fusion is influenced by previous work [4], which takes a pair of inputs consisting of a food image and text. In Figure 3, the left region represents the image model that uses image data as input. The right region represents the text model that uses text data as input. In the fusion method, both are then concatenated and fed into a neural network to produce the final prediction.

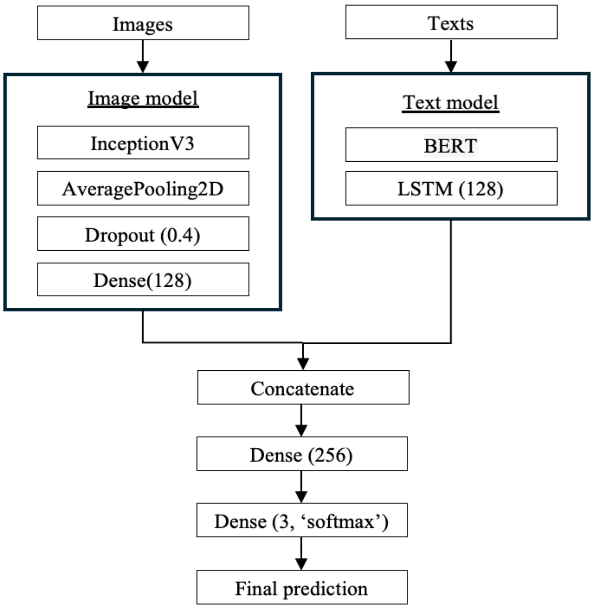

Figure 3: Diagram of the early fusion architecture.

This early fusion is learned to predict class probabilities independently for each sample. Therefore, the loss is calculated by Cross-Entropy Loss. We denote the output from concatenate from both the image and text encoders as Y_C, the image model as $\mathcal{H}$, the text model as $\mathcal{W}$, and the concatenate of the results from the image and text models as $\bigoplus$. The input image is defined as i, the input text is defined as t, and the ground truth is defined as G.

$$Y_C = \mathcal{H}(i) \oplus \mathcal{W}(t) \tag{3}$$

The cross-entropy loss function L is used as the objective function to measure the fitness between the output and target $(x,\ G)$. Let $x = softmax(Y_C)$ where c indicates the class label.

$$L(x, G) = -\sum_{i=1}^{c} G_i \log(x_i) \tag{4}$$

2.3.2 multimodal Siamese Neural Network

Siamese neural network is a class of neural network architectures that consist of two subnetworks with the same parameters and weights. This network accepts the distinct inputs that are either similar or dissimilar then they are joined by an energy function at the top, which computed distance metric between the highest-level feature representation on each side of the network [6].

Multimodal Siamese network extend traditional Siamese network to incorporate two different modalities. The proposed model compares two input samples (x_1, x_2), where $x_1 = (i_1, t_1)$ and $x_2 = (i_2, t_2)$ using the distance between them in the common space, calculated according to a contrastive loss function [7]. Here, i and t correspond to the food image and text, respectively. The first subnetwork processes the image and text from x_1, while the second subnetwork processes the image and text from x_2, which may belong to the same or opposite class as x_1.

Consider a multimodal dataset of sample pairs denoted as $\mathcal{D} = \{i_k, t_k\}_{k=1}^{N}$, where each i_k and t_k correspond to the food image and text of the k^{th} sample, respectively. The goal of this method is to develop two encoders that transform the image and text into a common space (ψ). The encoded outputs from both the image and text encoders are concatenated, which is used to measure the similarity between the embeddings of two sample pairs in the common space (ψ). The image encoder utilizes Convolutional Neural Network (CNNs) to map images into the common space, represented as $\mathcal{H}: \mathcal{I} \longrightarrow \psi$. The text encoder uses BERT and LSTM to project text inputs into the common space, denoted as $\mathcal{W}: \mathcal{T} \longrightarrow \psi$. Finally, the output from two subnetworks is passed through a contrastive loss function to generate the final output, as shown in Figure 4. [5,8].

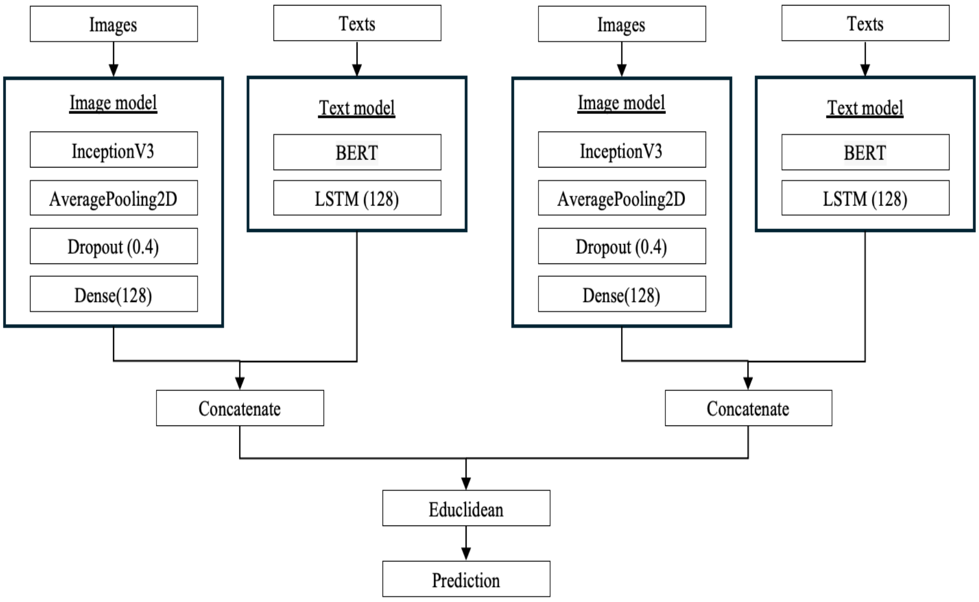

Figure 4: Diagram of the multimodal Siamese Neural Network architecture.

According to the contrastive loss defined in equation (5), U indicates whether the inputs (x_1, x_2) are similar or dissimilar. When U is equal to one, it indicates that the inputs (x_1, x_2) are similar representations, resulting in a small distance between them. Conversely, when U equal to zero, it indicates that the inputs (x_1, x_2) are dissimilar representations, resulting in a distance greater than a specified margin value (m). In this study, the margin m is set to 1.

$$L(D,U) = UD^2 + (1-U)\max(m-D,0)^2 \tag{5}$$

The output from first subnetwork is defined as Y_{C1} from input x_1, where $x_1 = (i_1, t_1)$

$$Y_{C1} = \mathcal{H}(i_1) \oplus \mathcal{W}(t_1) \tag{6}$$

The output from second subnetwork is defined as Y_{C2} from input x_2 where $x_2 = (i_2, t_2)$

$$Y_{C2} = \mathcal{H}(i_2) \oplus \mathcal{W}(t_2) \tag{7}$$

Here, D represents the Euclidean distance between concatenated outputs from two subnetworks, Y_{C1} and Y_{C2}.

$$D = \|Y_{C1} - Y_{C2}\|_2 \tag{8}$$

Since the contrastive loss measures the distance between inputs, this study uses it to classify inputs by selecting the pair with the smallest distance.

2.4 Evaluation

The primary metric used to evaluate model performance is accuracy. Additionally, precision and recall metrics are calculated.

3. Research Results and Discussion

This section evaluates the performance of early fusion and multimodal Siamese Neural Network in classifying food classes using text and image data. We first present results from the early fusion model (Section 3.1), followed by the results from the multimodal Siamese Neural Network under various configurations (Section 3.2). A comparative analysis of both models is then discussed in Section 3.3. Finally, Section 3.4 discusses the analysis of failure cases on larger database sizes.

3.1 Early fusion

The performance metrics of the early fusion model, applied to three food categories from the UPMCFood-101 dataset—bread pudding, chicken wings, and waffles—are displayed in Table 3. This table shows the precision and recall values for each class, revealing that bread pudding achieved the highest precision. Similarly, the highest recall rates were noted for waffles. The model demonstrated an overall accuracy of 0.960, which will serve as a benchmark for subsequent comparisons with multimodal Siamese Neural Network.

Table 3: Performance metrics of the early fusion model.

Class	Precision	Recall	Accuracy
Bread pudding	0.977	0.958	
Chicken wings	0.963	0.954	0.960
Waffles	0.942	0.970	

3.2 Multimodal Siamese Neural Network

The performance of the multimodal Siamese Neural Network (mSNN) was evaluated by varying the number of training pairs and adjusting the sample size per class in the testing database. The training was conducted with four distinct sets of pair counts: 7,200, 36,000, 72,000, and 108,000. In testing, each sample from the test set was paired with entries from a database to predict its class based on the lowest average of predictions. The number of samples per class in the database was varied, with configurations of 1, 10, 20, and 30. The results, as outlined in Tables 4(a) through 4(d), show significant improvements in precision, recall, and accuracy with increased training volumes and larger database sizes.

Initially, with 7,200 training pairs, both accuracy and other metrics like precision and recall were relatively low for most classes. However, as the number of training pairs increased to 36,000, there was a notable improvement across all metrics. The highest accuracy reached up to 0.97 with 72,000 training pairs, as detailed in Table 4(c), suggesting a correlation between the number of training samples and model accuracy. Conversely, the data from 108,000 training pairs, as shown in Table 4(d), exhibited consistent metrics across different testing configurations, potentially indicating overfitting; this implies that further increases in training data may not proportionally enhance performance.

The overall results demonstrate that increasing the amount of training data generally correlates positively with the accuracy of the model, highlighting the beneficial impact of larger datasets on model performance. However, excessive addition of the training dataset can lead to overfitting, where the model becomes too tailored to the training data and less effective at generalizing to new data. Furthermore, the results suggest that increasing the number of samples per class in the database does not enhance model accuracy. These findings and their implications are discussed in more detail in Section 3.4 of the document.

Table 4(a) Performance metrics of the mSNN model with 7,200 training pairs across different database sizes

	1 ITPC-DB			10 ITPC-DB			20 ITPC-DB			30 ITPC-DB		
Class	P	R	A	P	R	A	P	R	A	P	R	A
Bread pudding	0.533	0.354		0.000	0.000		0.000	0.000		0.000	0.000	
Chicken wings	0.451	0.840	0.518	0.346	1.000	0.390	0.091	0.005	0.337	0.322	1.000	0.322
Waffles	0.721	0.374		0.979	0.196		0.341	0.970		0.000	0.000	

Table 4(b) Performance metrics of the mSNN model with 36,000 training pairs across different database sizes

	1 ITPC-DB			10 ITPC-DB			20 ITPC-DB			30 ITPC-DB		
Class	P	R	A	P	R	A	P	R	A	P	R	A
Bread pudding	0.972	0.938		0.984	0.836		0.985	0.898		0.986	0.903	
Chicken wings	0.948	0.913	0.946	0.856	0.922	0.912	0.909	0.913	0.932	0.914	0.918	0.935
Waffles	0.920	0.983		0.909	0.974		0.909	0.983		0.913	0.983	

Table 4(c) Performance metrics of the mSNN model with 72,000 training pairs across different database sizes

	1 ITPC-DB			10 ITPC-DB			20 ITPC-DB			30 ITPC-DB		
Class	P	R	A	P	R	A	P	R	A	P	R	A
Bread pudding	0.982	0.982		0.982	0.956		0.982	0.978		0.974	0.982	
Chicken wings	0.991	0.959	**0.976**	0.991	0.959	**0.966**	0.991	0.959	**0.974**	0.991	0.959	**0.974**
Waffles	0.959	0.987		0.931	0.983		0.951	0.983		0.958	0.979	

Table 4(d) Performance metrics of the mSNN model with 108,000 training pairs across different database sizes

	1 ITPC-DB			10 ITPC-DB			20 ITPC-DB			30 ITPC-DB		
Class	P	R	A	P	R	A	P	R	A	P	R	A
Bread pudding	0.953	0.978		0.953	0.978		0.953	0.978		0.953	0.978	
Chicken wings	0.986	0.950	0.966	0.986	0.950	0.966	0.986	0.950	0.966	0.986	0.950	0.966
Waffles	0.962	0.970		0.962	0.970		0.962	0.970		0.962	0.970	

ITPC-DB: image text per class in database, P: precision, R: recall, A: accuracy

3.3 Comparative Analysis

To provide a meaningful comparison between the early fusion model and the multimodal Siamese Neural Network (mSNN), the configuration with 72,000 training pairs was chosen for the mSNN. This decision is based on the observation that the mSNN at 72,000 pairs achieved its highest overall accuracy, making it the most reliable and robust version of the model without showing signs of overfitting, which were observed with 108,000 pairs. When comparing the precision and recall across the three food classes—bread pudding, chicken wings, and waffles—the mSNN generally demonstrates higher precision and recall compared to the early fusion model. Specifically, the mSNN outperforms the early fusion model across all metrics when tested with 1 sample and 20 samples per class in the database.

In terms of accuracy, the mSNN with 72,000 pairs consistently outperforms the early fusion model across various database configurations (1, 10, 20, 30 ITPC-DB). The mSNN reaches a peak accuracy of 0.976 when tested with 1 database, while the early fusion model achieves an accuracy of 0.960, as shown in Table 5. Overall, the key takeaway from this comparison is that the mSNN with 72,000 pairs offers higher accuracy and slightly better performance on specific metrics compared to the early fusion model. These findings suggest that the mSNN is more suitable for applications where maximizing accuracy is critical, particularly when sufficient training data is available.

Although this study focuses on three specific food classes, the results indicate that the model architecture can be generalized to accommodate additional food classes. The structure of both the early fusion and mSNN models was not tailored specifically to these three classes, suggesting that the models are capable of handling other categories of food data. The ability to integrate both image and text data in a flexible manner makes the models applicable to a wide range of food types beyond those tested in this study.

Furthermore, the processing techniques used in this research, which fuse information from images and text, are designed to handle multimodal data inputs. This design allows the models to generalize effectively to different types of food classes without requiring adjustments to the model architecture. This suggests that, despite the limited number of classes tested, the findings can be extended to broader applications where a more comprehensive set of food categories is involved. The flexibility of the multimodal fusion approach enhances the potential for generalization to other domains or datasets with similar multimodal characteristics.

Table 5: Comparison of model accuracy

Model	Accuracy
Early fusion	0.960
mSNN (72,000 pairs, 1 ITPC-DB)	0.976
mSNN (72,000 pairs, 10 ITPC-DB)	0.966
mSNN (72,000 pairs, 20 ITPC-DB)	0.974
mSNN (72,000 pairs, 30 ITPC-DB)	0.974

3.4 Analysis of Failure Cases on Larger Database Sizes

While the multimodal Siamese Neural Network (mSNN) generally demonstrates strong performance, a significant drop in accuracy occurs as the number of database samples per class increases. This section explores the underlying causes of these failure cases and provides insights into the model's limitations.

Table 6 presents an example of misclassification, where a bread pudding image was incorrectly classified as waffles when tested with 10 samples per class in the database. This misclassification occurred due to the presence of a few outlier distances within the bread pudding class, such as 5.728 and 1.701, which disproportionately raised the average distance for the bread pudding class. The average distance calculated for bread pudding was 1.110, compared to 1.247 for chicken wings and 0.737 for waffles. These outliers caused the average distance for bread pudding to be higher than that for waffles, leading the model to incorrectly predict the test image as waffles, based on the lowest average distance.

This failure highlights a key vulnerability of the mSNN: its reliance on averaging distances can lead to errors when outliers are present. As the database size increases, the likelihood of such outliers also rises, leading to more frequent misclassifications. This analysis explains why the mSNN may perform better with smaller image text per class in databases but struggles with larger ones. To address this issue, future work could incorporate outlier detection mechanisms to improve the model’s robustness.

Table 6 Distance predictions of the mSNN for a test image of bread pudding against a database of 10 samples per class.

Database	Predict distance	Average distance per Class	Database	Predict distance	Average distance per Class	Database	Predict distance	Average Distance per Class
Bread pudding1	0.421	1.110	Chicken wings1	1.250	1.247	Waffles1	0.764	0.737
Bread pudding2	0.492		Chicken wings2	1.268		Waffles2	0.743	
Bread pudding3	0.482		Chicken wings3	1.254		Waffles3	0.722	
Bread pudding4	**5.728**		Chicken wings4	1.237		Waffles4	0.714	
Bread pudding5	0.414		Chicken wings5	1.252		Waffles5	0.762	
Bread pudding6	0.494		Chicken wings6	1.262		Waffles6	0.715	
Bread pudding7	0.488		Chicken wings7	1.178		Waffles7	0.685	

Table 6 (Continued)

Database	Predict distance	Average distance per Class	Database	Predict distance	Average distance per Class	Database	Predict distance	Average Distance per Class
Bread pudding8	**1.701**		Chicken wings8	1.253		Waffles8	0.749	
Bread pudding9	0.450		Chicken wings9	1.249		Waffles9	0.749	
Bread pudding10	0.428		Chicken wings10	1.263		Waffles10	0.764	

4. Conclusions and Recommendations

This study explored and compared two models for image and text data fusion in food classification: the early fusion model and the multimodal Siamese Neural Network (mSNN). The early fusion model demonstrated strong generalization capabilities, achieving an overall accuracy of 0.960. In contrast, the mSNN, when trained with 72,000 pairs, reached a peak accuracy of 0.976, outperforming the early fusion model in precision, recall, and accuracy when tested with smaller image text per class in databases. However, the mSNN exhibited vulnerabilities as database size increased, leading to reduced accuracy due to outlier effects that skewed average distance calculations. These findings suggest that while the mSNN offers superior accuracy in controlled and data-rich scenarios, its performance may degrade in larger, more varied datasets. In such environments, the early fusion model, though slightly less accurate, provides a more reliable and robust alternative.

Although this study focused on three specific food classes, the flexible architecture of both models suggests that they can be applied to other food categories without significant modifications. The results obtained here indicate that the models, particularly the mSNN, have the potential for generalization to a broader range of food classes in real-world applications. This suggests that, despite the limited number of classes tested, the models can be extended to include more diverse categories with minimal adjustments.

In future research, our objective is to address the issue of outliers in model predictions to improve the robustness of the mSNN, particularly when testing with larger databases. Additionally, further studies should evaluate the scalability of both models with significantly larger classes and diverse data types to better understand their limitations and identify potential enhancements for real-world applications. It is also recommended to incorporate a comparison with state-of-the-art models, such as Vision-based Transformers, which have demonstrated strong performance in visual tasks. This comparison could provide deeper insights into how transformer architectures might outperform traditional fusion models in the context of multimodal food classification.

5. Acknowledgements

I sincerely thank my advisors, Prof. Seksan Kiatsupaibul and Asst. Prof. Akarin Phaibulpanich, for their invaluable guidance. I am also deeply grateful to Miss Wassana Sintarasirikulchai for her encouragement and feedback, and to my family for their unwavering support.

6. References

[1] Bu L, Hu C, Zhang X. Recognition of food images based on transfer learning and ensemble learning. Plos one. 2024 Jan 19;19(1):e0296789.

[2] Mansouri M, Benabdellah Chaouni S, Jai Andaloussi S, Ouchetto O. Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: A systematic review. SN Computer Science. 2023 Jul 5;4(5):513.

[3] Abdulkadir Ş, Yaman A, Umit B. Food Image Classification with Deep Features. InComputer Science International Artificial Intelligence and Data Processing Symposium (IDAP) 2019.

[4] Gallo I, Ria G, Landro N, La Grassa R. 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); 2020 Nov 25. p. 1-6. IEEE. Available from: Kaggle https://www.kaggle.com/datasets/gianmarco96/upmcfood101.

[5] Chakladar DD, Kumar P, Roy PP, Dogra DP, Scheme E, Chang V. A multimodal-Siamese Neural Network (mSNN) for person verification using signatures and EEG. Information Fusion. 2021 Jul 1;71:17-27.

[6] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. InICML deep learning workshop 2015 Jul 6 (Vol. 2, No. 1, pp. 1-30).

[7] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) 2005 Jun 20 (Vol. 1, pp. 539-546). IEEE.

[8] Palazzo S, Spampinato C, Kavasidis I, Giordano D, Schmidt J, Shah M. Decoding brain representations by multimodal learning of neural activity and visual features. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020 May 20;43(11):3833-49.