A Multimodal Deep Learning Approach to Caloric Intake Estimation Using Visual, Physiological and Demographic Features

Afreen Ahmed
CSCE
Texas A&M University
College Station, TX
afreen04@tamu.edu

Rhea Sudheer
CSCE
Texas A&M University
College Station, TX
rheasudheer19@tamu.edu

Darshnilsinh Rana
CSCE
Texas A&M university
College Station, TX
darsh_04@tamu.edu

***Abstract*—Accurately tracking dietary intake is crucial for managing health conditions, improving nutrition, and developing personalized interventions for chronic diseases like diabetes and obesity. Traditional methods such as manual food logging or calorie estimation apps are often prone to inaccuracies and user fatigue, limiting their effectiveness. This challenge highlights the need for a more precise and automated method, leading to the design of a multimodal approach that integrates continuous glucose monitoring (CGM) data, meal photographs, and demographic information for improved caloric intake estimation.**

In this study, a model was developed combining convolutional neural networks (CNNs) for image analysis, with distinct encoders for each modality—CGM data, meal images, and demographic features. This integration of visual, physiological, and demographic data through advanced machine learning techniques enhances the model's accuracy, scalability, and ability to provide personalized insights. The proposed approach has the potential to revolutionize dietary assessment by delivering automated, real-time predictions and supporting healthcare providers in offering tailored nutritional recommendations for improved health outcomes.

***Index Terms*—Continuous Glucose Monitoring, Calorie Estimation, CNN, Pre-processing, Encoder, Multimodal model**

I. Introduction

Accurate nutrition estimation is essential for maintaining health, managing chronic diseases, and supporting overall well-being. Both overeating and undereating contribute to a range of health issues, making precise dietary assessment critical. Obesity and related conditions such as diabetes, cardiovascular disease, and metabolic syndrome are on the rise globally, while undereating can lead to malnutrition, weakened immune function, osteoporosis, and impaired organ function. Poor dietary intake is also associated with gastrointestinal disorders, nutrient deficiencies, and mental health issues such as depression and anxiety. Furthermore, it is not just total caloric intake that matters but also the balance of macronutrients—carbohydrates, proteins, and fats. An imbalanced diet can exacerbate health problems, with excessive carbohydrates causing glucose spikes and insufficient protein impairing muscle repair and overall metabolism. Ensuring a well-distributed intake of macronutrients tailored to individual needs is crucial for maintaining energy balance, promoting metabolic health, and preventing disease. Advanced tools for tracking both calorie and nutrient distribution are key to addressing the global challenges of diet-related diseases and malnutrition effectively. Historically, estimating calorie intake involved manual methods such as food weighing and consulting nutrition tables like the USDA Food Composition Database. While accurate, these methods were time-consuming and impractical for everyday use. The advent of mobile apps in the early 2000s marked a significant shift by streamlining calorie estimation through vast food databases. However, these tools still relied heavily on manual input and user estimations of portion sizes. Recent advancements in machine learning (ML) have revolutionized this field by enabling automated calorie estimation through computer vision techniques such as convolutional neural networks (CNNs). For instance, CNNs have been widely adopted for food recognition tasks due to their ability to analyze images of meals and estimate caloric content accurately. These systems leverage extensive nutritional databases to detect food items and assess portion sizes automatically. In addition to image-based approaches, wearable technologies like continuous glucose monitoring (CGM) devices provide critical insights into how food affects glycemic variability in real time. Studies have demonstrated that CGM data can be effectively analyzed using deep learning models such as recurrent neural networks (RNNs) or gated recurrent units (GRUs) to predict blood glucose levels [3]. Integrating CGM data with other modalities offers a more personalized approach to dietary assessment by accounting for individual physiological responses to meals. Multimodal learning frameworks that combine diverse data sources—such as meal photographs, CGM data, and demographic information—have shown promise in improving prediction accuracy by capturing complex interdependencies between these factors [1], [2]. This project aims to develop a multimodal approach to estimate lunch calorie intake by integrating CGM data, meal photographs, and demographic information. The dataset was collected from over 40 participants across several days, capturing detailed information about their meals, including photographs taken at breakfast and lunch times, glycemic variability data, and demographic details for each participant. By leveraging multimodal machine learning techniques that incorporate convolutional neural networks for image analysis alongside distinct encoders for each modality [3], [5], this approach seeks to enhance the accuracy of calorie estimation. The integration of visual, physiological, and demographic data enables a more comprehensive understanding of how meal composition and individual characteristics influence calorie intake. This work builds upon prior research in calorie estimation from food images [4], multimodal learning frameworks [1], personalized nutrition recommendations using machine learning [3], and advanced deep learning methods for health data analysis. Ultimately, this study aims to deliver automated real-time predictions that support healthcare providers in offering tailored nutritional recommendations for improved health outcomes. This version incorporates the requested citations seamlessly into the narrative while maintaining clarity and flow.

II. Dataset Description

A. Overview of Dataset Components

The dataset for this multimodal nutrition estimation project, which includes data from over 40 participants and spans a maximum of 9 days per participant, is comprised of several key components:

- Continuous Glucose Monitoring (CGM) Data
- Food Image Data
- Demographic and Viome (Microbiome) Data
- Meal calorie Labels

These components are organized into separate training and testing sets, with distinct CSV files for each modality.

B. Description of Each Modality

1) CGM Data: The CGM data provides real-time glucose readings collected at 5 minute intervals throughout the day. This data captures the dynamic glycemic profiles of participants in response to meals and other factors. CGM data can help infer calorie intake by detecting postprandial (after-meal) glucose spikes, which are typically larger after meals containing more carbohydrates or calories.

2) Food Image Data: The dataset includes photographs of breakfast and lunch meals consumed by participants during the period of the study. The meal photographs were captured during breakfast and lunch times, which are used to identify food items and estimate portion sizes, contributing to calorie and macronutrient estimation.

3) Demographic and Viome Data: This data includes participant-specific demographic information (such as age, gender, and diabetes status) as well as microbiome data, which may influence metabolic responses to meals, adding a personalized layer to the calorie estimation process.

4) Meal Calorie Labels: These serve as ground truth data for training and evaluating the model's accuracy in estimating caloric content of lunch meals, which serves as the target variable for the estimation model.

III. Methodology

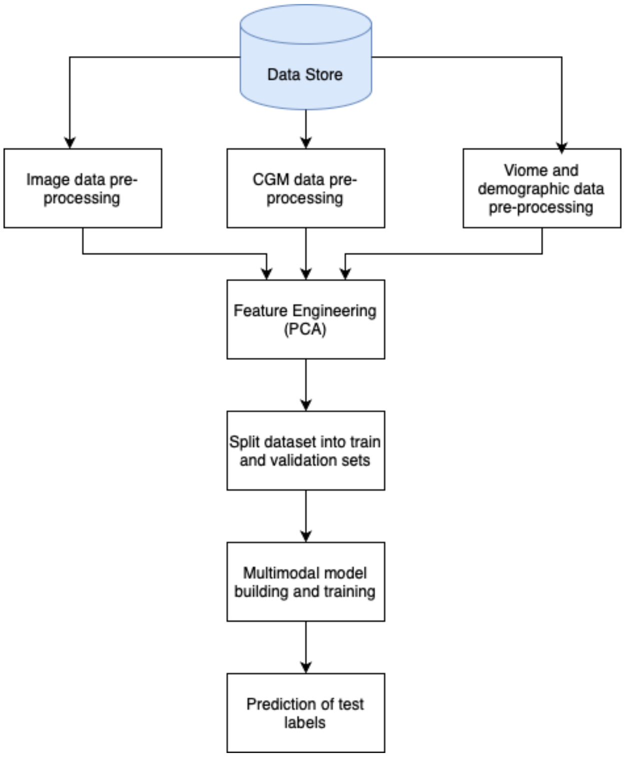
Data Store
Image data pre-processing
CGM data pre-processing
Viome and demographic data pre-processing
Feature Engineering (PCA)
Split dataset into train and validation sets
Multimodal model building and training
Prediction of test labels

Fig. 1. Overall System Architecture

This section outlines the methodology employed to prepare and integrate data from multiple modalities—image, Continuous Glucose Monitoring (CGM), and demographic-viome data—into a unified framework for model training. The pre-processing steps involve resizing and normalizing the image dataset, addressing missing CGM data using the MAGE algorithm, and transforming demographic and viome data into suitable formats for machine learning. Additionally, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the viome dataset, ensuring computational efficiency while retaining essential features. The model architecture, MultimodalNet, processes each modality through specialized encoders, followed by a fusion layer to combine the features for final prediction. The training process incorporates specific loss functions and optimizers to enhance model performance and prevent overfitting. The entire methodology is summarized in the flowchart shown in Figure 1, and the steps are explained in detail in the upcoming subsections.

A. Data Pre-Processing

1) Image Data Pre-Processing: To prepare the image dataset of meals taken during breakfast and lunch times for model training, several pre-processing steps were employed to ensure consistency and optimize model performance. Missing images were substituted with blank black images to maintain consistency in the dataset and avoid interruptions during model training.

The original images, which were of size 64x64x3 pixels as shown in Figure 2, were resized to a fixed 64x64 resolution to standardize input sizes across the entire dataset. This step is crucial as varying image dimensions can hinder computational efficiency and model performance. Additionally, the pixel values of the images were normalized to the range [0, 1], which is a standard practice to aid in faster convergence during training by preventing large gradient values that could destabilize the learning process.

Fig. 2. Sample Breakfast Image

2) CGM Data Preprocessing: The preprocessing of the Continuous Glucose Monitoring (CGM) data involved addressing several inconsistencies and preparing the dataset for model training. Initially, rows that contained no recorded CGM data were removed to ensure that only valid and complete data were used in the analysis.

To handle missing CGM data for breakfast and lunch times, the MAGE (Mean Absolute Glucose Excursion) algorithm was applied. MAGE is primarily used to quantify glucose variability, measuring the magnitude of excursions (i.e., increases or decreases) in glucose levels from baseline. While MAGE is typically used to assess glucose variability across a full time-series, in this case, it was adapted to predict missing CGM data during critical meal periods—specifically, breakfast and lunch. The algorithm estimates glucose fluctuations around these times by analyzing patterns in the rest of the CGM data, using the known data points to infer values for missing periods.

MAGE works by calculating the mean absolute differences between the peaks and valleys in a time-series, identifying periods of significant glucose variation. By applying this approach, missing CGM values during breakfast and lunch can be predicted based on the surrounding glucose fluctuations, ensuring continuity in the dataset and enabling more accurate modeling.

Once the missing data were addressed using MAGE, the CGM time-series sequences were padded with zeros to ensure consistent input length for model training.

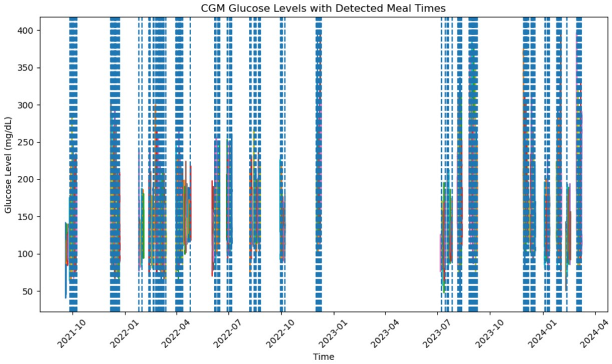

Fig. 3. CGM Glucose Levels with Detected Meal Times

3) Demographic and Viome Data Preprocessing: The Viome data initially consisted of a single column containing an array of comma-separated values. To make this data usable for machine learning models, the array was split into separate columns, each representing a specific feature or microbiome variable. `SimpleImputer` was used to handle missing values in numeric columns. Categorical variables such as Race and Diabetes status were present in the demographic data which were one-hot encoded to convert them into a format suitable for machine learning models.

B. Feature Engineering

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the numerical data of Viome Dataset, which initially contained 28 features, to a more manageable set of 22 features, while preserving as much variance as possible. By selecting the number of components that explained 90% of the variance, the reduced feature set retained the most important information, minimizing the loss of valuable data while improving computational efficiency.

C. Model Architecture

The MultimodalNet is designed to process and integrate information from multiple data modalities, including images, continuous glucose monitoring (CGM) data, and demo-viome data. This approach leverages the complementary information from each modality to enhance prediction accuracy. The network consists of several encoders that handle each modality independently, followed by a fusion layer that combines their feature representations for the final prediction.

1) Image Encoder: The image inputs, which represent breakfast and lunch meals, are processed through a convolutional neural network (CNN) to extract feature representations. This encoder utilizes two convolutional layers followed by max-pooling operations to reduce spatial dimensions while increasing the depth of the feature maps.

In the first layer, a 3-channel RGB image is convolved with 32 filters using a kernel size of 3 and padding of 1. This ensures that the output maintains the same spatial dimensions as the input image. After the convolution operation, a ReLU activation function is applied to introduce non-linearity, allowing the network to learn more complex patterns. A subsequent max-pooling layer with a stride of 2 is applied to downsample the feature map by a factor of 2, reducing the spatial dimensions while retaining the most important features.

The second convolutional layer takes the output from the previous layer, convolving it with 64 filters, again using a kernel size of 3 and padding of 1. This further increases the depth of the feature maps. Another max-pooling operation follows, reducing the spatial dimensions once more.

Following these convolutional and pooling operations, the output feature map is flattened into a 1D vector, which is then passed through a fully connected layer with 256 units. This layer serves to extract high-level features from the image, which are then used for downstream tasks. The final output of the image encoder is a 256-dimensional feature vector. Mathematically, the image encoding process can be represented as:

$$f_{\text{image}} = \text{ReLU}\left(\text{Linear}(\text{Flatten}(I), 256)\right)$$

where I denotes the input image, and f_{image} represents the feature vector extracted from the image.

2) CGM Encoder: The continuous glucose monitoring (CGM) data, assumed to have a length of 288 time steps, is processed by a fully connected neural network designed to extract relevant features from the sequential data. This encoder consists of two fully connected layers with 128 and 64 units, respectively.

The CGM data is first passed through a linear layer that projects the 288-dimensional input into a 128-dimensional feature space. A ReLU activation function is applied to introduce non-linearity. The output from this layer is then passed through a second linear layer, reducing the dimensionality to 64. The resulting feature vector, f_{cgm}, captures the most salient information from the CGM data.

The CGM encoding process can be represented mathematically as:

$$f_{\text{cgm}} = \text{ReLU}\left(\text{Linear}(c_{\text{gm}}, 128)\right)$$

$$f_{\text{cgm}} = \text{ReLU}\left(\text{Linear}(f_{\text{cgm}}, 64)\right)$$

where c_{gm} represents the input CGM data, and f_{cgm} is the output feature vector.

3) Demo-Viome Encoder: The demo-viome data, which encapsulates biological and microbiome-related information, is processed using another fully connected network. The input data, of size $d_{\text{demo\_viome}}$, is first projected into a 64-dimensional feature space, followed by a second fully connected layer that reduces the dimensionality to 32. This network serves to extract meaningful features from the demo-viome data, which are crucial for the downstream task.

The demo-viome encoding process is mathematically represented as:

$$f_{\text{demo\_viome}} = \text{ReLU}\left(\text{Linear}(demo\_viome, 64)\right)$$

$$f_{\text{demo\_viome}} = \text{ReLU}\left(\text{Linear}(f_{\text{demo\_viome}}, 32)\right)$$

where $demo\_viome$ represents the input demo-viome data, and $f_{\text{demo\_viome}}$ is the output feature vector.

4) Fusion Layer: Once the individual encodings for the images, CGM, and demo-viome data are obtained, they are concatenated into a single feature vector. This concatenation is done along the feature dimension (dimension 1). Specifically, the features from the breakfast image encoder, lunch image encoder, CGM encoder, and demo-viome encoder are concatenated as follows:

$$f_{\text{combined}} = [f_{\text{breakfast}}, f_{\text{lunch}}, f_{\text{cgm}}, f_{\text{demo\_viome}}]$$

This concatenated vector is then passed through a series of fully connected layers to fuse the individual modality features and generate a final prediction. The fusion layers consist of two linear transformations followed by ReLU activations, with the final output being a single scalar value. This fusion process can be represented as:

$$f_{\text{fusion}} = \text{ReLU}\left(\text{Linear}(f_{\text{combined}}, 128)\right)$$

$$f_{\text{fusion}} = \text{ReLU}\left(\text{Linear}(f_{\text{fusion}}, 64)\right)$$

$$\text{output} = \text{Linear}(f_{\text{fusion}}, 1)$$

5) Model Training: The model was trained for 50 epochs, allowing sufficient time for learning while preventing overfitting by avoiding excessive training. A batch size of 32 was chosen to balance training efficiency and memory constraints. The learning rate was set to 0.001, which is a commonly used value that facilitates stable and efficient convergence in most deep learning tasks, particularly when using optimizers like Adam.

- **Loss Function**: Root Mean Squared Relative Error (RMSRE) was selected as the loss function. RMSRE is effective for measuring the relative error between predicted and actual values, with an emphasis on minimizing the error in a way that accounts for differences in scale.
- **Optimizer**: The Adam optimizer was used due to its adaptive learning rate properties, which make it particularly suitable for training deep learning models with sparse gradients or noisy data.

IV. Results and Analysis

A. Model Performance Metrics

The model's performance was evaluated using predicted lunch calorie values, which were output as a single column of predictions corresponding to the test dataset. This test dataset included data from 9 subjects. The final training loss achieved by the model was 0.3339, indicating that the model had learned to predict the lunch calorie content with a reasonable degree of accuracy during the training process.

V. Discussion

A. Interpretation of Results

The multimodal neural network approach for estimating lunch calorie intake demonstrates promising results. The model's final training loss of 0.3339 using the Root Mean Squared Relative Error (RMSRE) indicates a moderate level of accuracy in predicting lunch calories. This suggests that the combination of CGM data, meal images, and demographic information can provide valuable insights into calorie estimation.

B. Limitations of the Current Approach

- Limited Dataset Size: The dataset consists of only 40 participants, each with data spanning up to 10 days. This small sample size may limit the model's ability to generalize to a larger, more diverse population. Smaller datasets often lead to overfitting.
- Potential Overfitting: The model's complex architecture, which integrates multiple data modalities (CGM data, meal photos, and demographic information), may exacerbate the risk of overfitting, particularly when trained on a limited dataset.
- Dependency on multiple data sources: The model relies on three distinct data sources: CGM data, meal photos, and demographic information. This creates a dependency on data availability and quality, which may not always be feasible in real-world scenarios. Ensuring the robustness of the model under varying data conditions will be crucial for real-world deployment.

C. Potential Applications and Implications

The model has significant potential applications in personalized nutrition and health management. By leveraging individual metabolic responses, it could assist in creating personalized nutrition plans, optimizing meal choices based on an individual's specific needs and preferences. Additionally, the model could play a crucial role in diabetes management by predicting calorie intake and its impact on glucose levels, offering valuable insights for better blood sugar regulation in diabetic patients. In weight management programs, accurate calorie estimation could enhance the effectiveness of strategies for both weight loss and weight gain, providing tailored recommendations based on precise nutritional data. Finally, the model could contribute to research in nutritional science by improving the understanding of how meal composition, individual characteristics, and calorie intake interact, ultimately advancing knowledge in dietary science and nutrition optimization.

VI. Conclusion

In conclusion, accurately tracking dietary intake is crucial for managing health conditions and improving nutrition, particularly for individuals with chronic diseases like diabetes and obesity. The development of the multimodal deep learning model to estimate lunch calorie intake involved several stages, from data preprocessing to model training and evaluation. The preprocessing phase addressed challenges like missing data, image resizing, and normalization. Using the MAGE algorithm to predict missing meal times based on CGM data was crucial for ensuring the model could process time-series data effectively. The model's architecture integrated multiple data sources, including images, CGM data, and demographic features, and was trained with a combination of hyperparameter tuning.

VII. Future Work

A. Suggestions for Improving the Model

Despite the promising performance of the model, several limitations hindered its generalization, including a small dataset and potential overfitting. Throughout the model training process, adjustments to dropout rates, batch sizes, and learning rates helped improve performance, but further fine-tuning is necessary. To enhance the robustness of the model, future work could incorporate advanced image processing techniques such as sophisticated image augmentation to increase data diversity and improve generalization. Additionally, leveraging time-series analysis techniques, such as recurrent neural networks (RNNs), could improve feature extraction from the CGM data. Rigorous feature selection for demographic and Viome data is also recommended to identify the most relevant variables for accurate calorie prediction.

Future improvements could also include the use of ensemble methods, transfer learning, and attention mechanisms to further boost performance. The journey of building this model revealed the complexities involved in integrating multimodal data sources and highlighted areas for optimization. This project offers valuable insights into the potential applications of deep learning in personalized nutrition and health management.

B. Optimization of Training Techniques

To enhance the model's performance and prevent overfitting, several advanced training techniques can be employed. First, incorporating cross-validation, such as k-fold cross-validation, would provide more robust performance estimates and help in fine-tuning the hyperparameters, ensuring the model generalizes well to unseen data. In addition, implementing learning rate scheduling strategies, such as cyclic learning rates or warm restarts, could help improve model convergence by adjusting the learning rate dynamically during training, thereby avoiding getting stuck in local minima. Lastly, regularization techniques like dropout or L1/L2 regularization should be applied to mitigate overfitting by reducing the complexity of the model and encouraging it to learn more generalizable features. Together, these techniques can lead to better optimization, enhanced generalization, and overall improved model performance.

References

[1] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. (2011). Multimodal Deep Learning. Proceedings of the 28th International Conference on Machine Learning (ICML).

[2] King, Ryan, Tianbao Yang, and Bobak J. Mortazavi. "Multimodal pretraining of medical time series and notes." In Machine Learning for Health (ML4H), pp. 244-255. PMLR, 2023

[3] Liu, Z., Li, J., Wu, Q., & Lee, Y. J. (2023). CalorieLLaVA: Image-based Calorie Estimation with Multimodal Large Language Models. Proceedings of the 2023 International Conference. Retrieved from UEC Japan publication archives.

[4] Tsolakidis, D., Gymnopoulos, L.P. and Dimitropoulos, K., 2024, August. Artificial Intelligence and Machine Learning Technologies for Personalized Nutrition: A Review. In Informatics (Vol. 11, No. 3, p. 62). MDPI.

[5] .Yang, S., Chen, M., Pomerleau, D., & Sukthankar, R. (2016). Calorie Estimation from Food Images Using Deep Learning. Proceedings of the ACM Multimedia Conference.

VIII. Work Division

Afreen Ahmed:

- Conducted thorough review of two papers to identify the most relevant studies for implementation.
- Research and experimentation with various pre-processing techniques for CMG data.
- Performed hypertuning of multimodal model to pick best hyperparameter values.
- Graphed key trends and findings.
- Documented abstract, introduction, dataset description, data pre-processing and result analysis in report

Rhea Sudheer:

- Conducted thorough review of two papers to identify the most relevant studies for implementation.
- Pre-processing techniques for Demographic and Viome data, along with dimensionality reduction.
- Implemented multimodal model
- Graphed key trends and findings.
- Documented overview of model architecture

Darshnilsinh Rana:

- Conducted thorough review of two papers to identify the most relevant studies for implementation.
- Pre-processing techniques for food image data
- Prepared test data for prediction
- Created the Powerpoint presentation to communicate our project findings.
- Helped with result analysis and discussion, conclusion and future work in the report.
- Compared different multimodal architectures to benchmark performance.