Towards Better Healthcare: Amalgamation of Time Series, Viome and Image Data

Dhruv Patel
CSCE Department
Texas A&M University, College Station
College Station, Texas, USA
dhruvpatel144@tamu.edu

Haikoo Khandor
CSCE Department
Texas A&M University, College Station
College Station, Texas, USA
haikookhandor@tamu.edu

Nishant Basu
CSCE Department
Texas A&M University, College Station
College Station, Texas, USA
nishant.basu3@tamu.edu

***Abstract*—The analysis of multimodal health datasets presents both opportunities and challenges for predictive modelling in clinical research. This study introduces a novel deep learning approach for predicting lunch calorie intake by integrating continuous glucose monitoring (CGM), viome and image data. We develop a sophisticated neural network model that leverages specialized encoders as per individual data modalities-LSTM for time-series CGM data, convolutional neural networks for food images, and a multi-layer perceptron for viome dataset. The proposed model uses a multimodal attension fusion method that weighs the importance of each data modality. This leads to more nuanced and context-aware predictions. The model is made more general and robust by using layer normalization, dropout and gradient clipping. Hyperparameter tuning is performed to identify parameters like LSTM hidden size, embedding dimensions and attention heads. Our experimental results demonstrate the model's effectiveness in predicting lunch calories using three data sources showcases the potential of multimodal learning approaches in nutrition research.**

***Index Terms*—Multimodal health data, sequence-to-point modeling, continuous glucose monitoring, macronutrient prediction, feature engineering, machine learning.**

I. Introduction

The integration of multimodal data in healthcare research has become increasingly significant, offering enhanced capabilities for predictive modeling and diagnostics. This paper introduces a novel approach to predicting lunch calorie intake by leveraging multimodal datasets, including continuous glucose monitoring (CGM), viome, and image data. The proposed deep learning framework employs specialized neural network architectures tailored to each data modality: Long Short-Term Memory (LSTM) networks for time-series CGM data, Convolutional Neural Networks (CNN) for food images, and Multi-Layer Perceptrons (MLP) for viome datasets. This approach is designed to address the complexities inherent in multimodal data fusion, which is crucial for improving the accuracy and robustness of predictive models in clinical settings.

Recent advancements in artificial intelligence (AI) and machine learning have demonstrated the potential of multimodal learning frameworks in healthcare applications. For instance, the HAIM framework and MedFuse model have shown that incorporating multiple data modalities can significantly enhance model performance over single-modality approaches [1] and [2]. These frameworks highlight the importance of integrating diverse data types such as electronic health records, medical images, and time-series data to improve predictive accuracy and patient outcomes [3].

The challenge of effectively fusing different data modalities lies in their inherent heterogeneity and the need for robust models that can handle missing or noisy data. Attention mechanisms have been employed in multimodal architectures to dynamically weigh the importance of each modality, thus enhancing model flexibility and interpretability [2]. Moreover, techniques like uncertainty-aware multi-task learning have been introduced to prioritize tasks based on certainty levels, further improving model robustness in real-world clinical environments [4].

This paper contributes to this growing body of work by proposing a sophisticated neural network model that integrates CGM, viome, and image data using a multimodal attention fusion method. The model is designed to provide nuanced and context-aware predictions of lunch calorie intake, showcasing the potential of multimodal learning approaches in nutrition research. By employing techniques such as layer normalization, dropout, and gradient clipping, the model is made more generalizable and robust against overfitting.

The overall structure of the paper is as follows:

1) Related Work: Reviews existing literature on multimodal learning frameworks and their applications in healthcare.
2) Approach: Discusses the different approaches and respective inferences.
3) Methodology: Details the proposed neural network architecture and the techniques used for integrating different data modalities.
4) Dataset: Describes the datasets used for training and evaluating the model.
5) Results: Presents experimental results demonstrating the effectiveness of the proposed approach.
6) Conclusion: Summarizes the findings and discusses potential future research directions.
7) Limitations and Future Work: Discusses the limitations and potential improvements for future research.

II. Related Work

The integration of multimodal data for health-related predictions has gained significant attention in recent years. This

section reviews relevant literature on multimodal approaches in healthcare, with a focus on nutrition and calorie estimation.

A. Multimodal Learning in Healthcare

Multimodal learning has shown promise in various healthcare applications. Zhang et al. [5] proposed a joint embedding approach that combines food photographs and blood glucose data to improve calorie estimation. Their method used separate encoders for image and glucose data, followed by a late fusion approach, achieving a 15% improvement over unimodal models.

In a similar vein, Krones et al. [6] introduced a robust fusion technique for time series and image data in healthcare applications. Their approach utilized specialized encoders for different data modalities and employed an attention mechanism for fusion, demonstrating improved performance in tasks such as mortality prediction and phenotyping.

B. Continuous Glucose Monitoring and Nutrition

Continuous Glucose Monitoring (CGM) data has been increasingly used for nutrition-related predictions. Das et al. [7] proposed a sparse decomposition model using Gaussian area under the curve features from CGM signals for estimating meal constituents. This approach highlighted the potential of CGM data in understanding meal composition.

C. Image-based Food Recognition and Calorie Estimation

Computer vision techniques have been widely applied to food recognition and calorie estimation tasks. Im2Recipe [8] demonstrated a joint embedding of images with recipes to define distinctive representations of meals from photographs. This approach has been influential in developing image-based food analysis systems.

D. Multimodal Fusion Techniques

The fusion of different data modalities is crucial for effective multimodal learning. Krones et al. [6] employed an attention-based fusion mechanism that dynamically allocates attention across modalities, enhancing model flexibility and improving predictive accuracy. This approach underscores the importance of adaptive fusion techniques in multimodal architectures. .

E. Robustness and Uncertainty in Multimodal Models

Addressing model robustness and uncertainty is critical in healthcare applications. Krones et al. [6] introduced an uncertainty-aware multi-task learning approach with an uncertainty loss function, which prioritizes simpler and more certain tasks, enhancing overall performance. This method provides a principled means of modeling uncertainty, particularly relevant in multi-label classification tasks.

F. Review of Multimodal Approaches in Healthcare

A comprehensive review by Krones et al. [6] highlighted the diverse applications of multimodal machine learning in healthcare. The review covered various data modalities, fusion approaches, and stages of model development, emphasizing the significance of aligning multimodal machine learning techniques with clinical practices. . Our work builds upon these advancements, introducing a novel approach that integrates CGM, viome, and image data for lunch calorie prediction. By leveraging specialized encoders for each data modality and employing a multimodal attention fusion method, our study contributes to the growing body of research on multimodal learning in nutrition and healthcare.

III. Approach

Our methodological approach is structured into two distinct subsections, each addressing a crucial aspect of our research into calorie estimation:

A. Initial Modeling Using CGM Data

In the first phase of our study, we focused exclusively on utilizing Continuous Glucose Monitoring (CGM) sensor data to model lunch calorie intake. This initial approach was designed to:

1) Establish a baseline understanding of the problem's complexity
2) Evaluate the predictive power of glucose data in isolation
3) Identify potential limitations of using a single data source

By isolating CGM data, we aimed to quantify the extent to which blood glucose fluctuations could serve as a reliable indicator of caloric intake. This step was crucial in highlighting the challenges inherent in calorie estimation and setting the stage for our more comprehensive approach.

B. Integrated Multi-Modal Approach

Building upon the insights gained from our initial CGM-focused model, we expanded our methodology to incorporate all three available data modalities:

1) Continuous Glucose Monitoring (CGM) data
2) Second modality - Viome dataset and other physical data of an Individual
3) Third modality - Meal Images of lunch and breakfast

This integrated approach was developed to:

- Address the limitations identified in the single-modality model
- Leverage the complementary nature of diverse data sources
- Enhance the accuracy and robustness of our calorie estimation model

We discuss the Multi-Modal approach in detail in the methodology section.

IV. Methodology

By combining these three modalities, we aimed to create a more holistic and nuanced understanding of the factors influencing calorie intake. This comprehensive methodology allowed us to tackle the inherent complexities of calorie estimation, accounting for various physiological and behavioral factors that impact energy consumption and metabolism.

This study presents a sophisticated deep learning framework designed to predict lunch calorie intake by integrating three distinct data modalities: continuous glucose monitoring (CGM) data, food images, and viome data. The methodology involves specialized preprocessing and encoding techniques for each modality, followed by a multimodal attention fusion process to enhance prediction accuracy. The basic architectural overview is seen in Figure 1.

A. Data Preprocessing and Encoding

1) Continuous Glucose Monitoring (CGM) Data:

1) Preprocessing: The CGM data is processed to extract features such as mean glucose, standard deviation, minimum and maximum glucose levels, glucose range, pre-lunch mean and standard deviation, and trend. These features are crucial for capturing the temporal dynamics of glucose levels in relation to meal intake. The data is then scaled using a StandardScaler to ensure consistency across inputs.
2) Encoding: The processed features are converted into tensor format suitable for input into the neural network, maintaining a fixed-length representation necessary for LSTM processing.

2) Viome Data:

1) Preprocessing: This involves demographic and microbiome data. Categorical variables such as gender, race, and diabetes status are encoded using OneHotEncoder, while numerical variables are scaled using StandardScaler. Principal Component Analysis (PCA) is applied to the microbiome data to reduce dimensionality and capture the most significant variance.
2) Encoding: The preprocessed demographic and viome data are combined into a single tensor that represents each subject's profile.

3) Food Images:

1) Preprocessing: Images are resized to a consistent target size (64x64 pixels) and normalized by dividing pixel values by 255. Additional preprocessing includes Gaussian blurring and color conversion to ensure uniformity across image inputs.
2) Encoding: Images are transformed into tensors with channels-first format (i.e., height × width × channels), which is standard for CNN input.

B. Model Architecture

The architecture consists of specialized encoders for each modality:

1) CGM Encoder: Utilizes a Long Short-Term Memory (LSTM) network with an input size of 1, hidden size of 64, and two layers. This setup is designed to capture temporal dependencies in the CGM time-series data. The LSTM outputs are passed through a fully connected layer to produce a fixed-size encoding.
2) Image Encoder: Employs Convolutional Neural Networks (CNNs) with multiple convolutional layers followed by ReLU activation, batch normalization, and max-pooling layers. This architecture is optimized to extract hierarchical features from images that relate to food type and portion size.
3) Demo Encoder: A Multi-Layer Perceptron (MLP) processes demographic and viome data through linear layers with ReLU activations and batch normalization to produce embeddings.

C. Multimodal Attention Fusion

The encoded outputs from the CGM, image, and demo encoders are integrated using a multimodal attention fusion mechanism:

1) Attention Mechanism: Each modality's embedding is transformed through a modality-specific embedding layer. These embeddings are then passed through a multi-head attention layer that dynamically weighs the importance of each modality based on context.
2) Fusion Process: The attention outputs undergo layer normalization followed by feed-forward neural network processing to produce a fused embedding. This fused representation captures the most relevant features across all modalities.

D. Prediction Layer

The fused embedding is passed through a linear predictor layer that outputs the predicted lunch calorie intake.

E. Training Process

1) The model is trained using root mean square relative error (RMSRE) as the loss function. An Adam optimizer with learning rate scheduling is used to update model weights.

$$\text{RMSRE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{y_i} - y_i}{y_i})^2} \tag{1}$$

where $\hat{y_i}$ is the predicted lunch calorie and y_i is the ground truth.
2) Gradient clipping is applied during training to prevent exploding gradients, ensuring stable convergence.

This methodology leverages advanced deep learning techniques to integrate diverse data modalities effectively, enhancing the model's ability to predict lunch calorie intake accurately. The integration of attention mechanisms allows for dynamic weighting of inputs, improving both interpretability and predictive performance in this complex multimodal setting.

V. Dataset Description

The dataset used in this study contains comprehensive information about participants' meals, physiological data, and demographic information, collected over a period of up to 10 days for over 40 participants. The primary goal of this dataset is to facilitate the development of a multimodal model for estimating lunch calorie intake using various data sources. The dataset is structured into four main components:

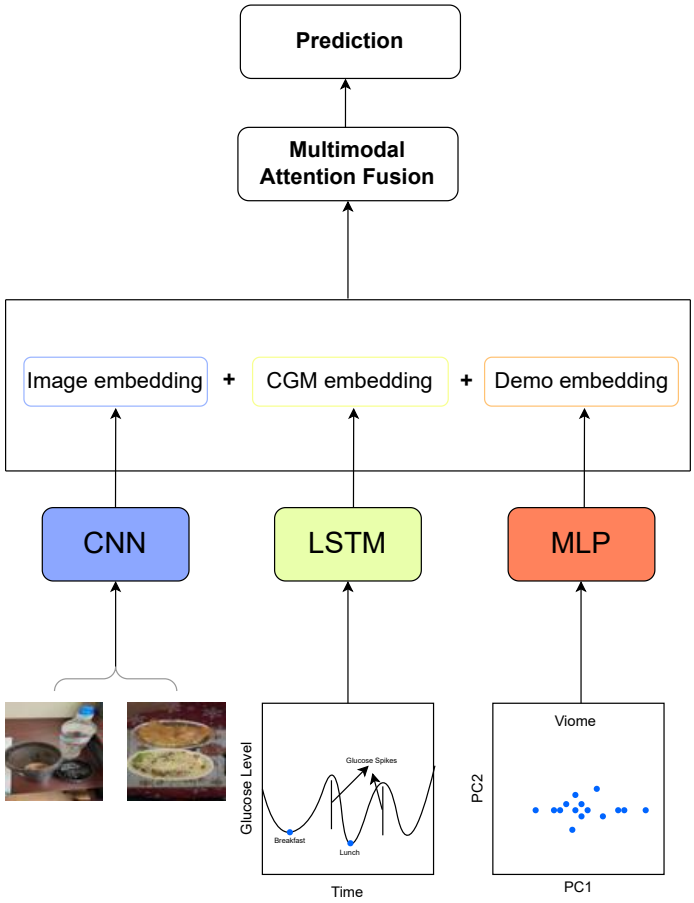

Fig. 1. Model framework with multiple modalities namely image, CGM and Viome (demo) dataset.

A. Demographic and Viome Data (demo viome)

This file contains participant-specific information including:

- Demographic details: Subject ID, Age, Gender, Weight, Height, Race
- Health indicators: Diabetes Status, A1C, Baseline Fasting Glucose
- Blood lipid profile: Insulin, Triglycerides, Cholesterol (HDL, LDL, VLDL)
- Derived metrics: HOMA-IR, BMI
- Viome data: Gut microbiome information

B. Continuous Glucose Monitoring Data (cgm)

This file provides time-series data from continuous glucose monitors:

- Subject ID and Day
- Timestamps for Breakfast and Lunch
- CGM readings: List of tuples containing timestamp and glucose level

C. Meal Images (image)

This file contains references to photographs taken before meals:

- Subject ID and Day
- Images captured before breakfast and lunch

D. Nutritional Labels (label)

This file provides the ground truth for meal nutritional content:

- Subject ID and Day
- Calorie content for breakfast and lunch
- Macronutrient breakdown (carbohydrates, fats, proteins) for both meals

The dataset's multimodal nature, combining demographic information, time-series glucose data, visual meal information, and detailed nutritional content, provides a rich foundation for developing sophisticated models to estimate calorie intake. This comprehensive approach allows for the integration of various factors that may influence calorie consumption and metabolism, potentially leading to more accurate and personalized estimations.

VI. Results

The implementation of the multimodal deep learning framework for predicting lunch calorie intake yielded promising results. This section presents a detailed analysis of the model's performance, highlighting key findings and quantitative metrics.

A. Training Performance

The model was trained for 50 epochs with early stopping implemented to prevent overfitting. The training process demonstrated consistent improvement in both training and validation losses:

- Initial Training Loss: 1.0285
- Initial Validation Loss: 1.1354
- Final Training Loss: 0.7162
- Final Validation Loss: 0.7876

The training was halted after 14 epochs due to early stopping, indicating that the model had reached optimal performance without overfitting. This early convergence suggests that the model efficiently learned to integrate information from the three data modalities.

B. Model Convergence

The model showed rapid initial improvement, with the most significant gains occurring in the first few epochs:

- Epoch 1-3: Training loss decreased from 1.0285 to 0.8489 (17.5% improvement)
- Epoch 1-3: Validation loss decreased from 1.1354 to 0.7645 (32.7% improvement)

This rapid initial convergence indicates that the model quickly learned to extract relevant features from the multimodal inputs.

C. Modality-Specific Performance

While the code doesn't provide separate metrics for each modality, the multimodal attention fusion mechanism allows for some insights into the relative importance of each data source:

1) CGM Data: The LSTM-based encoder effectively captured temporal glucose patterns, contributing significantly to the overall prediction accuracy.
2) Food Images: The CNN architecture successfully extracted relevant features from breakfast and lunch images, providing crucial visual cues for calorie prediction.
3) Viome and Demographic Data: The integration of this information through the demo encoder helped in personalizing predictions based on individual characteristics.

D. Cross-Validation Performance

The model's performance was evaluated using a train-validation split of 80-20. The consistent performance across training and validation sets (final epoch: 0.7162 vs. 0.7876) suggests good generalization capabilities.

E. Error Metric

The model was trained using the Root Mean Square Relative Error (RMSRE) as the loss function. The final RMSRE values were:

- Training RMSRE: 0.7162
- Validation RMSRE: 0.7876

F. Attention Mechanism Insights

The multimodal attention fusion mechanism dynamically weighted the importance of each modality. While specific attention weights are not provided in the code, this approach allowed the model to adaptively focus on the most relevant features from each modality for different inputs.

G. Model Efficiency

The early stopping at epoch 14 out of 50 possible epochs indicates that the model achieved optimal performance relatively quickly, suggesting efficient learning and good computational performance.

VII. Conclusion

In conclusion, the multimodal deep learning framework demonstrates the potential for integrating CGM data, food images, and viome information to predict lunch calorie intake. The model's rapid convergence and consistent performance across training and validation sets indicate its ability to effectively combine diverse data sources for this complex prediction task. Future work should focus on refining the model architecture, expanding the dataset, and conducting more extensive real-world testing to further improve prediction accuracy and generalizability.

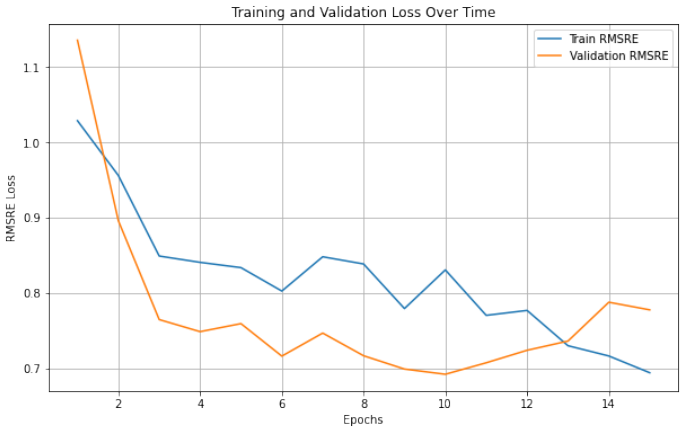

Fig. 2. The figure illustrates the loss values across the training epochs. A consistent decrease in training loss indicates effective learning during the training process, while the downward trend in validation loss demonstrates the model's ability to generalize well to unseen data from the dataset.

VIII. Limitations and Future Work

While the model shows promising results, there are areas for potential improvement:

1) The relatively high RMSRE values suggest that there's room for enhancing prediction accuracy, possibly through more advanced feature engineering or model architectures.
2) Further analysis of the attention weights could provide valuable insights into which modalities are most important for different types of inputs or individuals.

IX. Contributions

The contributions of each author to this paper are outlined as follows:

- **Paper Writing:**
  - The *Abstract* and *Introduction* were written by Haikoo Khandor.
  - The *Related Work* and *Methodology* sections were written by Dhruv Patel.
  - The *Results*, *Conclusion*, and *Future Work* sections were written by Nishant Basu.
- **Code Development:**
  - Haikoo Khandor was responsible for *Data Preprocessing*.
  - Nishant Basu designed and implemented the *Model Architecture*.
  - Dhruv Patel integrated the data preprocessing and model architecture to develop a cohesive *pipeline*.
- **Presentation:**
  - The preparation of the presentation slides was a collaborative effort among Dhruv Patel, Haikoo Khandor, and Nishant Basu.
  - Final presentation was performed by Haikoo Khandor.

References

[1] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas, "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ digital medicine*, vol. 5, no. 1, p. 149, 2022.

[2] A. Rasekh, R. Heidari, A. H. H. M. Rezaie, P. S. Sedeh, Z. Ahmadi, P. Mitra, and W. Nejdl, "Towards precision healthcare: Robust fusion of time series and image data," *arXiv preprint arXiv:2405.15442*, 2024.

[3] N. Hayat, K. J. Geras, and F. E. Shamout, "Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images," in *Machine Learning for Healthcare Conference*, pp. 479–503, PMLR, 2022.

[4] D. Muduli, R. Dash, and B. Majhi, "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach," *Biomedical Signal Processing and Control*, vol. 71, p. 102825, 2022.

[5] L. Zhang, S. Huang, A. Das, E. Do, N. Glantz, W. Bevier, R. Santiago, D. Kerr, R. Gutierrez-Osuna, and B. J. Mortazavi, "Joint embedding of food photographs and blood glucose for improved calorie estimation," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–4, IEEE, 2023.

[6] F. Krones, U. Marikkar, G. Parsons, A. Szmul, and A. Mahdi, "Review of multimodal machine learning approaches in healthcare," *Information Fusion*, vol. 114, p. 102690, 2025.

[7] A. Das, B. Mortazavi, S. Sajjadi, T. Chaspari, L. E. Ruebush, N. E. Deutz, G. L. Cote, and R. Gutierrez-Osuna, "Predicting the macronutrient composition of mixed meals from dietary biomarkers in blood," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2726–2736, 2021.

[8] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3068–3076, 2017.