Joint Embedding of Food Photographs and Blood Glucose for Improved Calorie Estimation

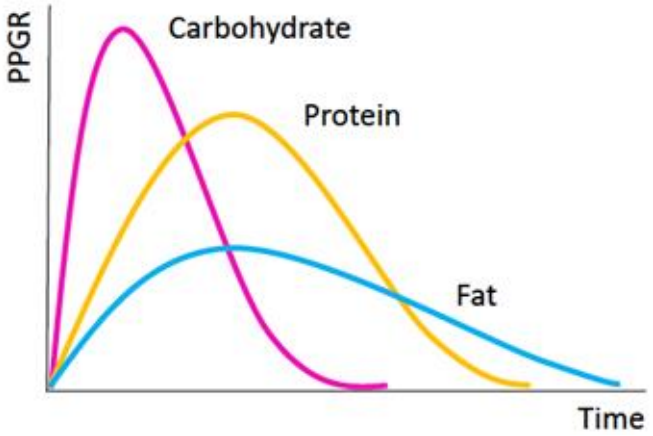
BHI 2023
Cited by 14

Lida Zhang, Sicong Huang, Anurag Das, Edmund Do, Namino Glantz, Wendy Bevier, Rony Santiago, David Kerr, Ricardo Gutierrez-Osuna, and Bobak J. Mortazavi

Introduction

Motivation: Why Combine CGM and Food Images?

- Diet monitoring is important for preventing and managing Type 2 diabetes.
- CGMs can estimate meal information, especially carbohydrates.
- Food images can estimate nutrition information such as calories.



Problem with CGM-only

- The same meal can create different glucose responses in different people.
- Different meals can also create similar glucose patterns.

Problem with image-only

- Photos may miss hidden ingredients like oil, sugar, sauces, or cooking style.

Paper's solution: Combine CGM + food images using late fusion to improve calorie estimation.

METHODOLOGY

133-dimensional fused vector → fully connected layer(s) + ReLU

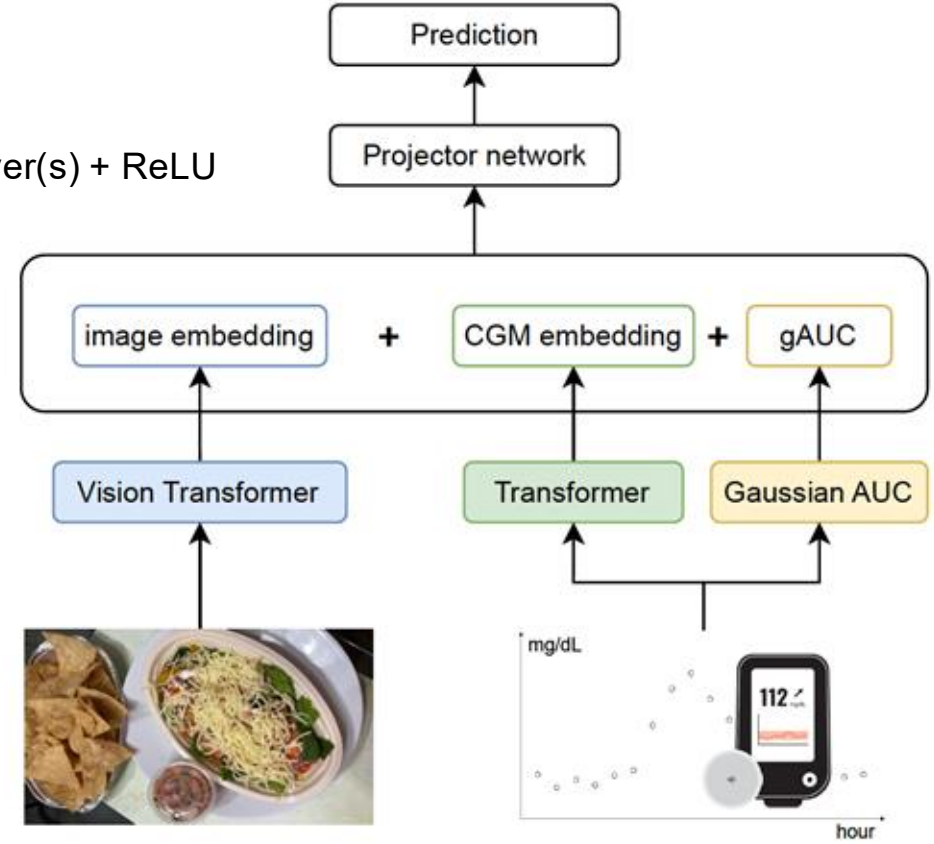
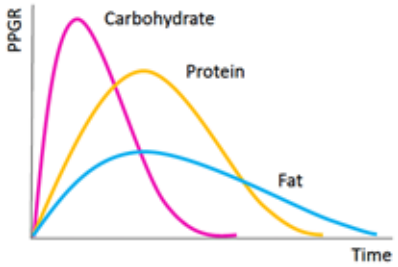


Fig. 1. The model framework of macronutrient prediction with multiple modalities of data from CGMs and food images.

Gaussian AUC

$$gAUC_k \approx \int_0^{180} x(t) G_k(t)\, dt$$

$$[gAUC_1, gAUC_2, gAUC_3, gAUC_4, gAUC_5]$$



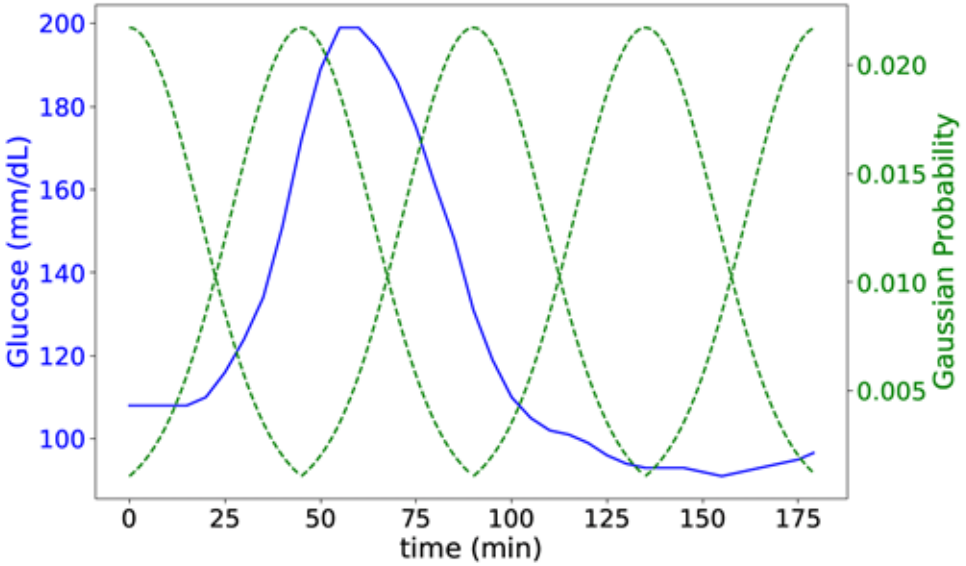


Fig. 2. An example of five Gaussian kernels.

CGM Pipeline: Step-by-Step

1. **Raw CGM Input**

   The Freestyle Libre Pro records interstitial glucose every 15 minutes. For a 3-hour post-meal window, that's ~13 data points per meal.

2. **Linear Interpolation → 1/min**

   Resample to 1-minute resolution → 180 data points. This makes the signal dense enough for a Transformer to learn temporal patterns.

3. **gAUC Feature Extraction (5 Kernels)**

   Five Gaussian functions of different widths are convolved with the glucose curve. Each captures the area under the smoothed signal. Result: 5 scalar features summarizing the shape of the glucose response (peak height, width, spread).

4. **Transformer Encoder**

   The 180-point time series is fed into a Transformer encoder (4 self-attention heads, 4 stacked layers, hidden size 64). Self-attention lets the model weight earlier vs. later time points dynamically — no need to hand-engineer 'peak at 45 min'. Output: a 64-dimensional CGM embedding.

5. **CGM Representation = Embedding + gAUC**

   The 64-dim transformer embedding is concatenated with the 5 gAUC features → a 69-dimensional vector representing the full CGM signal.

Image Pipeline: Step-by-Step

1. **Raw Food Photo**

   Participant photos taken at meal start and end. Variable sizes and aspect ratios.

2. **Resize to 112 × 112 pixels**

   Standardized for the ViT model. Smaller than typical (224×224) to save compute, but sufficient for this dataset scale.

3. **Patch Embedding (Vision Transformer)**

   ViT slices the image into 784 non-overlapping 4×4 pixel patches. Each patch is flattened into a vector. This is the ViT's way of treating an image like a sequence — similar to how NLP transformers treat a sentence as a sequence of word tokens.

4. **Transformer Encoder on Patches**

   The sequence of patch embeddings is processed by a Transformer encoder (2 self-attention heads, stacked layers, dropout 0.2). Self-attention across patches lets the model relate, e.g., the protein-rich portion of the plate to the carb-heavy portion. Output: a 64-dimensional image embedding.

5. **Image Representation = 64-dim Vector**

   This compact vector summarizes the visual content of the meal — colors, textures, portions — that correlate with caloric content.

Predictive Task & Evaluation Metrics

Metric 1: NRMSE

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2}$$

- Measures **relative calorie prediction error**
- Lower NRMSE = more accurate calorie prediction

Metric 2: Pearson Correlation

$$r = corr(y, \hat{y})$$

- Compares true calories vs. predicted calories
- Checks whether predictions follow the same **low-to-high calorie pattern**
- Higher correlation = better separation of low- and high-calorie meals

Dataset

Study Design

- 27 participants (10 healthy, 12 pre-diabetes, 5 T2D)
- 10-day protocol
- Abbott Freestyle Libre Pro CGM — reads every 15 min
- Fixed breakfasts (shakes) + fixed lunches (Chipotle)
- Participants photographed meals before & after eating
- 3 hours of post-prandial glucose captured per meal
- 20 unique meals total (10 breakfast, 10 lunch)

TABLE I

COMPOSITION OF MEALS AND THE CODE OF LOW (L) OR HIGH (H) MACRONUTRIENTS FOR CARBOHYDRATES, PROTEINS, FATS, AND FIBERS.

Breakfast Meal (BM)			Lunch Meal (LM)		
Index	Description	Calorie	Index	Description	Calorie
BM1	LLLL	268	LM1	HHHH	1180
BM2	HLLL	448	LM2	HLHL	830
BM3	HHLL	608	LM3	LHLL	435
BM4	HLHL	712	LM4	HLLL	555
BM5	HHHH	902	LM5	LLLL	355
BM6	LLLL	268	LM6	HHHH	1180
BM7	HLLL	448	LM7	HLHL	830
BM8	HHLL	608	LM8	LHLL	435
BM9	HLHL	712	LM9	HLLL	555
BM10	HHHH	902	LM10	LLLL	355

Results

Ablation Study

- **CGM-only**
- **Image-only**
- **CGM + Image fusion**

CGM-only Models

- Linear Regression, XGBoost: use **gAUC features**
- LSTM, Transformer: use **interpolated CGM time-series**

Image-only Models

- VGG16, VGG19, ResNet18, ResNet50, ViT

CGM + Image Fusion

- CGM encoder embedding + **gAUC features**
- Image encoder embedding
- Late fusion through a fully connected projector network

Training Setup

- 60% train / 20% validation / 20% test
- 10 repeated experiments
- Loss function: minimize **NRMSE**

Data	Model	NRMSE	Correlation
CGM-only	Linear Regression	0.72 (0.03)	0.24 (0.02)
	XGBoost	0.52 (0.02)	0.42 (0.03)
	LSTM	0.41 (0.03)	0.34 (0.02)
	Transformer	0.40 (0.04)	0.40 (0.03)
Image-only	VGG16	0.42 (0.04)	0.23 (0.03)
	VGG19	0.43 (0.02)	0.20 (0.04)
	ResNet18	0.42 (0.03)	0.31 (0.02)
	ResNet50	0.41 (0.03)	0.30 (0.01)
	ViT	0.43 (0.02)	0.22 (0.03)
CGM-image	LSTM-VGG16	0.36 (0.03)	0.38 (0.02)
	LSTM-VGG19	0.39 (0.02)	0.29 (0.03)
	LSTM-ResNet18	0.40 (0.01)	0.31 (0.03)
	LSTM-ResNet50	0.39 (0.02)	0.36 (0.04)
	LSTM-ViT	0.35 (0.02)	0.46 (0.02)
	Transformer-VGG16	0.36 (0.03)	0.34 (0.02)
	Transformer-VGG19	0.38 (0.02)	0.33 (0.04)
	Transformer-ResNet18	0.40 (0.04)	0.28 (0.04)
	Transformer-ResNet50	0.39 (0.01)	0.36 (0.03)
	Transformer-ViT	**0.34 (0.01)**	**0.52 (0.02)**

Limitations & Takeaways

⚠ **Small cohort:**	27 subjects, 20 meal types — needs larger-scale validation
⚠ **Subject variability:**	Healthy vs. pre-diabetic vs. T2D all pooled — individual metabolic differences drive noise
⚠ **Controlled meals only:**	Fixed Chipotle lunches / protein shakes — may not generalize to real-world diverse meals
⚠ **Calorie only:**	Macronutrients (carbs, fats, protein) not predicted separately — next step

Questions?

Thanks for your attention