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

Lida Zhang[1], Sicong Huang[1], Anurag Das[1], Edmund Do[1], Namino Glantz[2], Wendy Bevier[3], Rony Santiago[3], David Kerr[4], Ricardo Gutierrez-Osuna[1], and Bobak J. Mortazavi[1]

Computer Science & Engineering, Texas A&M University, College Station, TX[1], Santa Barbara County Education Office, Santa Barbara, CA[2], Sansum Diabetes Research Institute, Santa Barbara, CA[3], Diabetes Technology Society, Burlingame, CA[4]

{lidazhang, siconghuang, anuragdiisc.ac.in, rgutier, bobakm}@tamu.edu,
nglantz@sbceo.org, wbevier@sansum.org, rofsanti@ucsc.edu,
kerr@diabetestechnology.org

Abstract—Type 2 diabetes has a significant impact on **indviduals**' health and well-being, and diet monitoring is an important tool in treating individuals. Accurate estimation of meal intake is essential for promoting diet and behavior interventions. While continuous glucose models (CGMs) have demonstrated the ability to estimate carbohydrate quantities in meals, CGMs-alone have been insufficient in capturing other meal nutritional information due to the different types of food and people's health conditions. Therefore, we propose a multi-modality model for augmenting CGM-based inverse metabolic models by using both CGMcaptured interstitial glucose data and food image data. A late fusion approach is used to aggregate the extracted glucose information from the attention-based Transformer and Gaussian area under the curve (gAUC) features, and image information from the vision transformer. In this study, we build calorie estimation models on **20 meals with 30 participants**, with meals with known fixed caloric content. Our joint embedded approach to learning calorie estimations from both CGM data and image data achieved an average Normalized Root Mean Squared Error (NRMSE) of 0.34 for calorie prediction, with a correlation of 0.52, a 15.0% improvement over CGM-only models and 17.1% over image-only models.

Index Terms—Machine learning, Continuous Glucose Monitors, Diabetes, Diet monitoring, Nutrition

I. Introduction

Type 2 Diabetes (T2D) is a major condition that leads to over one million deaths in 2020 in the United States alone [1]. A primary component to prevent the progression from pre-diabetes to T2D is monitoring and controlling diet. With the development of remote sensing and machine learning analytics, automated techniques have explored aiding diet monitoring. These techniques include nutrition estimation of meals from both remote sensing and computer vision-based models. Continuous glucose monitors (CGMs), for example, have been used to estimate the constituents of a meal, potentially resulting in tools to log food intake automatically [2], such as estimating carbohydrate quantity [3], [4] and even estimating proteins and fats when paired with other sensors [4]. Computer vision techniques, which process photographs of meals, have also been used to estimate meal nutritional content, such as calories [5]–[8].

This work was supported, in part, by the National Science Foundation under awards IIS #2014475 and the PATHS-UP Engineering Research Center under National Science Foundation Award # 1648451.

However, macronutrient predictions from either CGMs or images have some limitations. Predicting meal macronutrients from post-prandial glucose responses (PPGR) is a complex many-to-one inverse mapping [2], [9], with individual variability leading to significant changes in PPGR from the same meals across participants [4], [10]. This variability in absorption and metabolism limits CGM-based model efficacy in macronutrient estimation. Similarly, additives (i.e., sugars and salts) are a source of uncertainty for food image-based models, where even cooking style can greatly impact nutrition estimation. As a result, many computer vision studies are limited to a certain type of food [11], such as Chinese food [12], Thai food [13] and Indonesian food [14]. The two modalities, however, may provide complementary information. Therefore, we design a multi-sensor fusion approach and multimodal modeling using both CGMs and food images.

We demonstrate our multimodal model improvement on the estimation of calories in a meal. A late fusion architecture [15] is applied to create a joint embedding of information from CGM and images and we demonstrate the improvement in calorie estimation from the joint embedding rather than learning from each individual modality. We conduct an ablation study on models that estimate calories in meals from just CGMs, from just images, and then jointly from both modalities, from a set of known breakfasts and lunches in a cohort of 27 participants.

II. Related Work

Recent work has explored the potential of using PPGRs from CGM devices for meal monitoring purposes. Das et al. proposed a sparse decomposition model using Gaussian area under the curve (gAUC) features from PPGR signals for the same purpose [9]. We leverage these gAUC features for modeling both as baseline comparisons and in our joint architecture.

In addition, we compare several computer vision techniques for identifying meal information. IM2Recipe demonstrated a

joint embedding of images with recipes to define distinctive representations of meals from photographs [5]. We adopt this

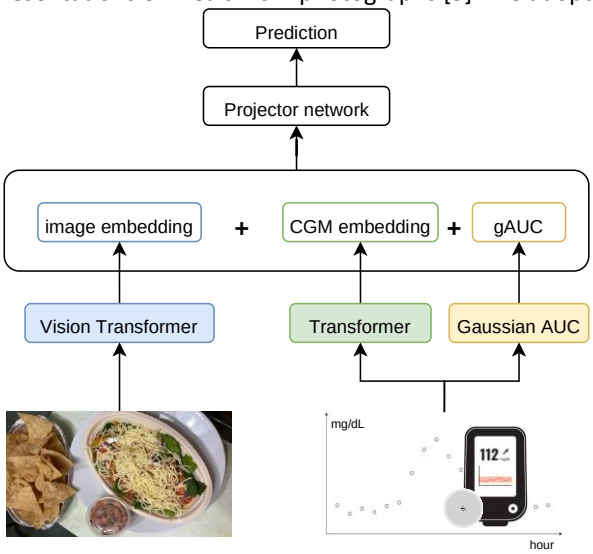

Fig. 1. The model framework of macronutrient prediction with multiple modalities of data from CGMs and food images.

approach of modeling food images into an embedded space for joint learning, using vision transformers as our primary image pretrained backbone [16]. In addition, we compare our technique against a series of baseline computer vision backbones, pretrained to be agnostic of downstream predictive tasks [17], [18].

III. Methodology

Our multimodal model is illustrated in Fig 1. We developed modality-specific models that contain three feature extractors that extract image embedding from food images, CGM embedding from CGM readings and gAUC features from CGM readings. We then concatenated all embeddings with a latefusion mechanism through a fully-connected projector network to generate calorie estimations. This section describes the data processing, modeling, and evaluation approach taken. The study will be described further in Section IV. In brief, each participant (N=27) in our study captured a photograph of each breakfast and lunch, with known calories and macronutrient composition, and consumed them while also wearing a CGM, in a 10-day study.

A. Data Preprocessing and Feature Extraction

The CGM sensor used in this study recorded interstitial glucose levels every fifteen minutes. We applied a linear interpolation to process the CGM data to have a frequency of every minute. After interpolating the data, we extracted the gAUC features from prior work [9]. Specifically, we applied five Gaussian-based kernel functions to extract the features. Each Gaussian kernel was convolved with the time series data, which results in smoothed signal, and then each smoothed signal was used to obtain the statistical gAUC value by calculating the total area under the smoothed signal curve. Figure 2 is an example of this process with five gAUC kernels.

For image data, we resized all the images to be the same size. Standard size is commonly used in deep neural network

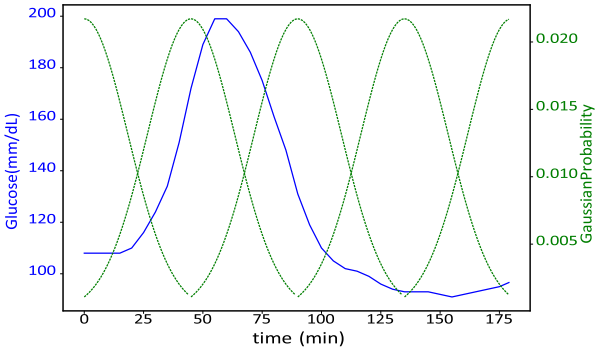

Fig. 2. An example of five Gaussian kernels.

models, which require input images to be of a fixed size. Facing the challenge of limited computation resources, we reduced the standard image size to 112 by 112 pixels. Reducing the image size can also help model training by reducing the impact of variations in image size and aspect ratio.

B. Calorie Estimation Modeling

For time-series CGM data, we applied an attention-based Transformer [19] to process the data. Since we were building a supervised learning task, instead of using the original encoder-decoder architecture of Transformer for the purpose of data reconstruction, we only utilized the encoder part for data representation learning and remove the decoder part for reconstruction. Transformers use a multi-head self-attention mechanism that allows them to learn important temporal relationships and dependencies within time-series data, which is crucial for accurate predictions. We designed our Transformer architecture of four self-attention stacks with four heads and a fully-connected layer with a hidden size 64, and obtain a final CGM embedding size of 64. A dropout layer with a

dropout rate of 0.2 is also applied in each stack to avoid overfitting. By attending to different time points in the input sequence, attention-based transformer models can effectively capture patterns and trends in the CGM data that may be representative of different macronutrients (and calories) in a meal. We then took that embedded data and fused it with five more nodes, representing the gAUC features manually extracted, to represent the CGM data.

We use vision transformers (ViT) on the images of food, leveraging its ability to learn hierarchical representations of image features to effectively analyze food images and predict the macro-nutritional content of different types of food. ViT can learn different information from food images, such as their

color, texture, and shape, which provides important information about their nutritional content. In brief, ViT breaks down an image of 112 by 112 into 784 patches with a patch size of 4, and then flattens and embeds them into a sequence of vectors. These patch embeddings are then fed into a standard Transformer encoder with a stacked self-attention with two heads, a feed-forward layer, and a dropout layer with a dropout rate of 0.2 to present a final image embedding of 64 size. For additional detail, we refer you to [16].

Finally, we used a late fusion approach, similar to that from Shukla et al. on time-series and text data [15], to aggregate the embedding information into a single, joint model. We designed a fully-connected projector network for a more comprehensive representation of the data and then make the final predictions from the output of this network. The late fusion approach allows each modality of data to be processed independently, which can result in more accurate and robust embeddings. In our study, CGM and images may have very different structures and features; processing them separately allows each modality to be optimized for its own unique characteristics, and the concatenation of them effectively leverages the complementary information from each modality.

C. Predictive Tasks and Evaluation Metrics

The model was trained to optimize calorie estimation of each meal, as its regression task. To evaluate the performance of the model, two common metrics used were the Normalized Root Mean Squared Error (NRMSE) and Pearson's Correlation Coefficient (correlation). The NMRSE is defined as:

$$NRMSE = \sqrt{\frac{1}{n}(\frac{y-\bar{y}}{\bar{y}})^2}$$

where n is the number of samples, y is the predicted value, and $\bar{y}$ represents the ground truth. NRMSE measures the average error between the predicted and true values, normalized by the true value, while correlation measures the strength and direction of the linear relationship between the predicted and true values. NRMSE is particularly useful in evaluating the accuracy of the model in terms of relative errors, which is important when dealing with nutritional data that can vary widely in magnitude (e.g. quantity of carbohydrates versus quantity of fats in a meal). A lower NRMSE indicates that the model is able to predict calorie values with greater accuracy and precision.

IV. Experiments and Results

A. Dataset

We collected meal images and CGM data in a trial conducted with 27 participants (Advarra IRB Pro00049227). Of the 27 participants, 12 had pre-diabetes, 10 were considered healthy, and the remaining had T2D. Each participant wore an Abbott Freestyle Libre Pro sensor on an arm to capture glucose data. The Abbott Freestyle Libre Pro is a blinded CGM that captures interstitial glucose readings every 10 minutes. Then, each day over a ten-day protocol, participants were given a breakfast shake (with known calories and macronutrient composition), asked not to consume anything for three hours, given a lunch from Chipotle (with known calories and macronutrient composition), asked not to consume anything else for the next three hours, then were given free choice of dinners. For each meal, participants were asked to take an image of the meal at the start of each meal, and a photograph afterwards to show the end of the meal and whether food remained. In addition, for the dinner, they were asked to provide myfitnesspal logs of the meal. In this study, we use the three hours post-prandial data

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

from breakfasts and lunches, whose description and calories are shown in Table I.

B. Experiment Setup and Hyperparameter Tuning

We designed our experimental setup to conduct ablation studies across modality-specific models, and then compared a variety of fusion models. We compared the model performance using CGM data only, image data only, and then we compared models using both CGM and image data, intended to understand if the multiple modalities of data actually benefit the modeling. For CGM data, we compared against [9] by using the gAUC features to implement a generalized linear regression and tree-based XGBoost. In addition, we evaluated two deep learning models LSTM and Transformer. For image data, five state-of-the-art deep learning models were compared: VGG16, VGG19, Resnet18, Resnet50, and ViT. For the multiple modalities of data, we tested different combinations of the two deep learning models for CGM and all five models for image data, given the fusion of the models with the predictions from the other CGM models is not possible for joint embedding and backpropagation of the loss function.

Hyperparameter tuning is applied to all the models, including {dropout rate of 0-0.2, batch size of 8-128, learning rate of $1e^{-2}$-$1e^{-4}$, hidden size of 64-512, weight decay of 0-0.2}. We also applied the activation function ReLU for the projector

layers of late fusion. We selected the best model and ran ten repeated experiments for each model. In each experiment, we shuffled all the meals, and randomly selected 60% data for training, 20% for validation, and 20 % for testing. The loss function used sought to minimize the NRMSE of calorie estimation. The mean NRMSE and its standard deviation were calculated based on all ten experiments for each model.

C. Result and Analysis

Table II shows the results of our experiment, and our best result of an NRMSE of 0.34 for calorie estimation. We observe that using both CGM and image data has a significant improvement over single modality calorie estimation. Our proposed model using Transformers and ViT improves the performance of calorie prediction by 10.8% compared to the best CGM model, and 19.5% to the best image model, to an NRMSE of 0.34.

TABLE II
MACRONUTRIENT PREDICTION PERFORMANCE COMPARISON AMONG DIFFERENT DATA MODALITIES AND MODELS

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)

When focusing on model selection, we observe that, for CGM data, both LSTM and transformer perform much better than linear regression and XGBoost. This result shows better robustness of the deep learning models than traditional machine learning models, and the variation among subjects could be the reason for the low performance of traditional machine learning models. The five image models do not make a significant difference in calorie prediction; however, ViT provides for stronger joint embedding.

V. LIMITATION AND FUTURE WORK

In this study, we build calorie estimation models on 20 meals with 27 participants. Ultimately, although calorie estimation is improved, there is room for further improvement in both NRMSE and correlation for the model, as well as extending the estimation to individual macronutrients. In addition, the variation among subjects (across healthy to those with diabetes) is likely to be a factor challenging the modeling. We plan to identify and account for additional sources of subjectspecific variation and incorporate that into future IMMs.

VI. CONCLUSION

In this study, we propose a calorie prediction IMM using both CGM and food image data. A transformer was used for CGM data extraction, and vision transformer was applied for image data. All the features are aggregated through a projector using the late fusion mechanism. The experimental results show that our calorie IMM with multiple modalities of data has significant improvement over models with a single data modality using three hours of post-prandial CGM data. Further, our proposed model outperforms all the baseline models and demonstrates that multiple ”views” of meals are needed for accurate diet monitoring technologies.

REFERENCES

[1] S. L. Murphy, K. D. Kochanek, J. Xu, and E. Arias, “Mortality in the united states, 2020,” 2021.

[2] B. J. Mortazavi and R. Gutierrez-Osuna, “A review of digital innovations for diet monitoring and precision nutrition,” *Journal of diabetes science and technology*, vol. 17, no. 1, pp. 217–223, 2023.

[3] S. Sajjadi, A. Das, R. Gutierrez-Osuna, T. Chaspari, P. Paromita, L. E. Ruebush, N. E. Deutz, and B. J. Mortazavi, “Towards the development of subject-independent inverse metabolic models,” in *ICASSP 20212021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3970–3974.

[4] 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.

[5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3020–3028.

[6] S. Mezgec, T. Eftimov, T. Bucher, and B. K. Seljak, “Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment,” *Public health nutrition*, vol. 22, no. 7, pp. 1193–1202, 2019.

[7] K. Motoki, T. Saito, S. Suzuki, and M. Sugiura, “Evaluation of energy density and macronutrients after extremely brief time exposure,” *Appetite*, vol. 162, p. 105143, 2021.

[8] M. B. Gillingham, Z. Li, R. W. Beck, P. Calhoun, J. R. Castle, M. Clements, E. Dassau, F. J. Doyle, R. L. Gal, P. Jacobs *et al.*, “Assessing mealtime macronutrient content: patient perceptions versus expert analyses via a novel phone app,” *Diabetes technology & therapeutics*, vol. 23, no. 2, pp. 85–94, 2021.

[9] A. Das, S. Sajjadi, B. Mortazavi, T. Chaspari, P. Paromita, L. Ruebush, N. Deutz, and R. Gutierrez-Osuna, “A sparse coding approach to automatic diet monitoring with continuous glucose monitors,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2900–2904.

[10] A. Das, B. Mortazavi, N. Deutz, and R. Gutierrez-Osuna, “Modeling individual differences in food metabolism through alternating least squares,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 2988–2992.

[11] Y. Lu, T. Stathopoulou, M. F. Vasiloglou, L. F. Pinault, C. Kiley, E. K. Spanakis, and S. Mougiakakou, “gofoodtm: an artificial intelligence system for dietary assessment,” *Sensors*, vol. 20, no. 15, p. 4283, 2020.

[12] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, “Chinesefoodnet: A large-scale image dataset for chinese food recognition,” *arXiv preprint arXiv:1705.02743*, 2017.

[13] N. Hnoohom and S. Yuenyong, “Thai fast food image classification using deep learning,” in *2018 International ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI-NCON)*. IEEE, 2018, pp. 116–119.

[14] S. Giovany, A. Putra, A. S. Hariawan, L. A. Wulandhari, and E. Irwansyah, “Indonesian food image recognition using convolutional neural network,” in *Artificial Intelligence Methods in Intelligent Algorithms: Proceedings of 8th Computer Science On-line Conference 2019, Vol. 2 8*. Springer, 2019, pp. 208–217.

[15] S. N. Shukla and B. M. Marlin, “Integrating physiological time series and clinical notes with deep learning for improved icu mortality prediction,” *arXiv preprint arXiv:2003.11059*, 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.